Igeta, Takahashi, and Matsui (2018)

STAT 778 Project

Tom Wallace

George Mason University

Spring 2018

Table of Contents

1. Background
2. Methods
3. Implementation in C
4. Simulation Study
5. Conclusion

What is overdispersion?

Occurs when a dataset exhibits higher variance than expected under the assumed distribution

Commonly occurs with **count data**

Motivating example: Poisson

- Mean $= \lambda$
- Var $= \lambda$
- Cannot change λ to correct for variance without skewing mean

What are the consequences of overdispersion?

If not corrected, overdispersion results in distorted test statistics and estimated standard errors

Practical consequences can be severe: count data is common in clinical trials

- E.g., anti-epilepsy drug might use *number of seizures over study period* as measure of drug efficacy
- Failure to correct for overdispersion could cause Type II error (failure to adopt beneficial treatment) or, even worse, Type I error (adopt treatment that is not beneficial and may even be harmful)

How can we correct for overdispersion?

Correct root cause of overdispersion if possible (e.g., zero-inflated models)

Use count distribution that allows adjustment of variance independent from mean (e.g., negative binomial)

Approach of authors:

- Want to estimate treatment effect in Poisson model $\lambda_i = \exp\left(\beta_0 + \beta_1 X_{ij}\right)$, but data is over-dispersed
- Distinguish between *true* variance function V (unknown) and *working* variance function $\tilde{V}$
- Chances are low that we guess perfectly such that $V = \tilde{V}$

Goal of paper

Test statistics, power, and sample size calculations that are robust to misspecification of the variance function ($V \neq \tilde{V}$)

Table of Contents

1. Background
2. Methods
3. Implementation in C
4. Simulation Study
5. Conclusion

Original contributions

Test statistic

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

Power of test using that statistic

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

Sample size required to achieve desired power

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

How to use them to determine sample size

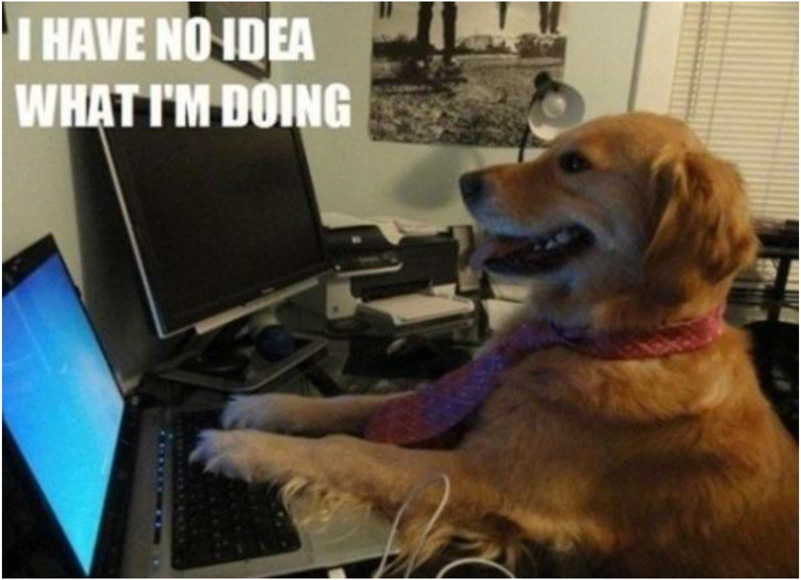
Authors propose the following procedure

1. Consider multiple true variance functions, pick most plausible (how?)
2. Specify working variance function
3. Conduct sensitivity analysis of power under different misspecifications
4. Adopt sample size achieving desired power

Table of Contents

1. Background
2. Methods
3. Implementation in C
4. Simulation Study
5. Conclusion

How I felt reading and trying to replicate this paper



My progress

Unable to replicate paper

Primary fault lies with my own limitations

- Paper assumes that readers are familiar are with many things not actually covered in paper
- This probably is a fair assumption for a peer-reviewed article, but I was out of my depth

That said. . . this was a poorly written paper

- Complex and inconsistently applied notation (e.g., $\hat{\tilde{\phi}}_p^*$)
- Apparent errors
- Confusing presentation

I will briefly cover some stumbling blocks

Simultaneous estimation

To calculate Z, we need to estimate $\hat{\tilde{\phi}}_p^*$, the estimated dispersion parameter of the working variance function under the null hypothesis

To estimate $\hat{\tilde{\phi}}_p^*$, we need $\hat{\beta_0}^*$, the estimated “base rate” parameter under the null hypothesis

To estimate $\hat{\beta_0}^*$, we need $\hat{\tilde{\phi}}_p^*$

Summary: we need A to calculate B, and B to calculate A

Solution: simultaneous numerical estimation of both, which apparently is a common approach in GEE

Estimation of treatment effect

We need to estimate $\hat{\beta}_1$ via quasi maximum likelihood (QML)

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

$\tilde{V}$ takes $\hat{\tilde{\phi}}_p^*$ as a parameter, which I failed to find (per previous slide)

I identified a working variance function $\tilde{V}$ for which $\hat{\tilde{\phi}}_p^*$ is irrelevant in our QML equations

Used Newton-Raphson to estimate, but failed to converge (singular matrix)

Unclear whether error is in my math or in my coding

Table of Contents

1. Background
2. Methods
3. Implementation in C
4. Simulation Study
5. Conclusion

Could not actually conduct study, but path is clear

Simulation study

1. Assume some true variance function and some true treatment effect $(\exp(\beta_1))$
2. Use authors' methods to calculate required sample size
3. Generate simulated data for control and treatment group
4. Conduct test of difference between groups using authors' Z-statistic and some assumed working variance function
5. Record whether Type I or Type II error occurred
6. Repeat many times, calculate empirical Type I and Type II error rate

If we do this for many cases in which $V \neq \tilde{V}$, and the empirical results match our theoretical expectations, we conclude that the authors' claims of robustness to misspecification of variance are supported

Authors' results

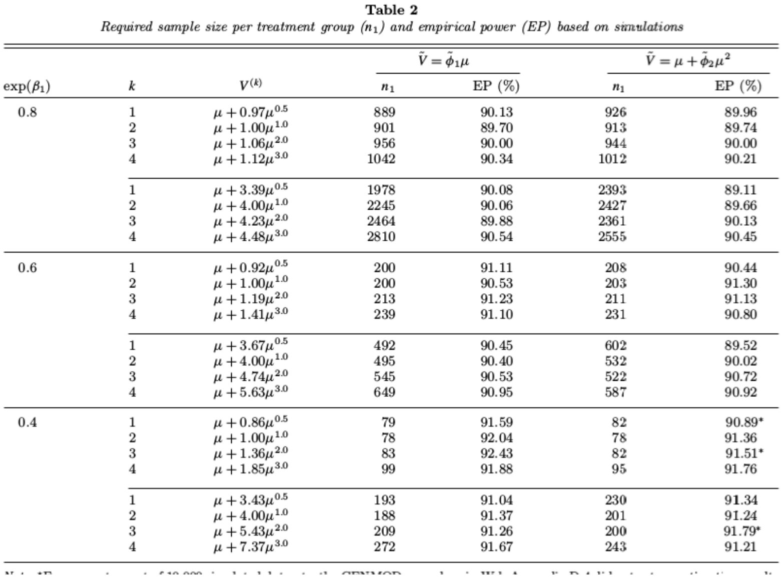


Table 2

Required sample size per treatment group (n_1) and empirical power (EP) based on simulations

			$\tilde{V} = \tilde{\phi}_1\mu$		$\tilde{V} = \mu + \tilde{\phi}_2\mu^2$	
$\exp(\beta_1)$	k	$V^{(k)}$	n_1	EP (%)	n_1	EP (%)
0.8	1	$\mu + 0.97\mu^{0.5}$	889	90.13	926	89.96
	2	$\mu + 1.00\mu^{1.0}$	901	89.70	913	89.74
	3	$\mu + 1.06\mu^{2.0}$	956	90.00	944	90.00
	4	$\mu + 1.12\mu^{3.0}$	1042	90.34	1012	90.21
	1	$\mu + 3.39\mu^{0.5}$	1978	90.08	2393	89.11
	2	$\mu + 4.00\mu^{1.0}$	2245	90.06	2427	89.66
	3	$\mu + 4.23\mu^{2.0}$	2464	89.88	2361	90.13
	4	$\mu + 4.48\mu^{3.0}$	2810	90.54	2555	90.45
0.6	1	$\mu + 0.92\mu^{0.5}$	200	91.11	208	90.44
	2	$\mu + 1.00\mu^{1.0}$	200	90.53	203	91.30
	3	$\mu + 1.19\mu^{2.0}$	213	91.23	211	91.13
	4	$\mu + 1.41\mu^{3.0}$	239	91.10	231	90.80
	1	$\mu + 3.67\mu^{0.5}$	492	90.45	602	89.52
	2	$\mu + 4.00\mu^{1.0}$	495	90.40	532	90.02
	3	$\mu + 4.74\mu^{2.0}$	545	90.53	522	90.72
	4	$\mu + 5.63\mu^{3.0}$	649	90.95	587	90.92
0.4	1	$\mu + 0.86\mu^{0.5}$	79	91.59	82	90.89*
	2	$\mu + 1.00\mu^{1.0}$	78	92.04	78	91.36
	3	$\mu + 1.36\mu^{2.0}$	83	92.43	82	91.51*
	4	$\mu + 1.85\mu^{3.0}$	99	91.88	95	91.76
	1	$\mu + 3.43\mu^{0.5}$	193	91.04	230	91.34
	2	$\mu + 4.00\mu^{1.0}$	188	91.37	201	91.24
	3	$\mu + 5.43\mu^{2.0}$	209	91.26	200	91.79*
	4	$\mu + 7.37\mu^{3.0}$	272	91.67	243	91.21

Table of Contents

An educational experience

This paper clearly is intended to be practical and useful

But, poor quality of presentation means the target audience (working statisticians) will ignore it—too much effort to decipher, too little time

It doesn't have to be this way—many of the most-cited statistics papers ever are easy to read (e.g. Efron 1979)

- This is part of why they're so highly cited!

There is an important lesson here: **effective communication of your methods is just as important as formal correctness**