STAT 0017: Selected Topics in Statistics **2022/23**

Written Report

Submission Date: March 3, 2023 *Author: Alden & James*

Report

We were provided clinical trial data testing the effect of progabide as an anti-epileptic drug. This data includes the patient's age, baseline seizure counts, treatment status, and biweekly seizure counts (or "counts") during the treatment period.

Exploratory Analysis

We begin by visualising scatter plots of counts comparing periods, where each ppoint is determined by the data from a single patient.

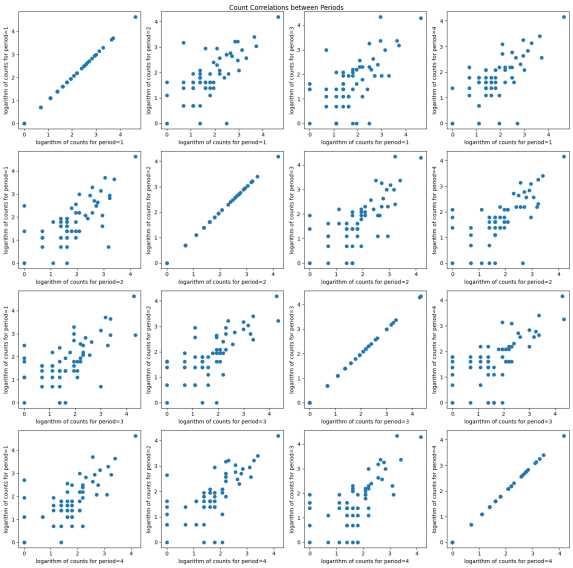

Figure 1.1. Scatter plots of seizure counts comparing different periods

We can see that there are correlations for seizure counts within the periods of each patient. This will be taken into account during our model choice and data simulation.

We can also observe the histograms for baseline and age data. These helped to inform data simulation choices from a qualitative perspective.

To study the effects of progabide on seizure occurrence in epilepsy patients, we choose to build a poisson regression model containing various covariates on the response variable: biweekly seizure counts in epilepsy patients.

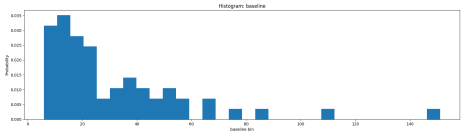

Figure 1.2. Histogram for baseline data

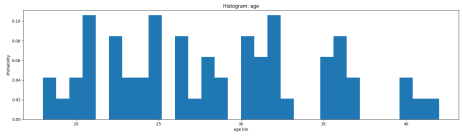

Figure 1.3. Histogram for age data

Covariate Selection

Our first model covariate is treatment status because the anti-epilectic effects of progabide is the main hypothesis under investigation. If the hypothesis is true, it is expected to reduce the count in treatment groups compared to placebo groups.

The next two covariates are period and treatment by period, where period indicates the number of fortnights since the trial started. Including period allows us to account for possible placebo effects in patients and including treatment by period allows the overall effect of progabide to be partitioned.

Finally, to adjust for individual backgrounds, we include the age and baseline seizure count. Age is included since there exists clinical evidence that age can affect an individual's number of seizures [1, 2]. Baseline seizure counts are included since treatment reduction effects might not be detected if the effect is small and many individuals have high baseline counts.

Model Specification

After covariate selection, we build the model:

$$log(\mu_{ij}) \;=\; \text{Treatment}_i + \text{Period}_j + (\text{Treatment} * \text{Period})_j + \text{Age}_i + \text{Baseline}_i \quad (1)$$

for $i \in \{1, ..., 59\}$ and $j \in \{1, 2, 3, 4\}$ where μ_{ij} is the expected count for the i^{th} individual at time period j.

The correlated nature of counts for each patient from period one to four requires us to adjust for underlying covariance structure in the trial data. This is done using the Generalised Estimating Equation (GEE) framework, provided via the statsmodels Generalized Estimating Equation package in Python.

Model (1) was fitted with the poisson distribution family parameter sor integer counts and an exchangeable correlation structure. This correlation structure was used since the counts for a patient are expected to be correlated in a constant, patient-specific way, while each patient is independent from each other. Hence, the correlation structure is expected to be fixed within-patient while varying between patients.

Results & Inference

The GEE output of coefficient estimates is provided in Table 1:

Table 1.1. Table 1: Table of Model 1 coefficient estimates

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	0.641511	0.342253	1.874377	6.087847e-02	-0.029292	1.312315
age	0.023466	0.011706	2.004615	4.500424e-02	0.000523	0.046409
treatment	-0.075089	0.184442	-0.407115	6.839236e-01	-0.436589	0.286411
baseline	0.022750	0.001241	18.328286	4.921674e-75	0.020317	0.025183
period	-0.042805	0.059059	-0.724788	4.685818e-01	-0.158559	0.072948
treatment:period	-0.031472	0.068448	-0.459790	6.456672e-01	-0.165628	0.102684

The model also provides an estimate of the scale/dispersion: 5.11 and within-patient correlation: 0.401.

From Table 1, we can identify the effect of the treatment covariate on counts. Specifically, we see that the estimated coefficient of the treatment covariate is -0.0751, and treatment by period covariate has a coefficient of -0.0315.

For two patients with the same age and baseline and different treatment status, at week 1 of the trial, the placebo-treated patient will have an expected seizure count of X, the patient undergoing progabide treatment will have an expected biweekly seizure count of $e^{-0.0751+(1)(-0.0315)}X = 0.896X$. Here, progabide treatment seems to reduce expected seizure count by a factor of 0.896.

While we observe that progabide treatment induces an effect on the expected seizure counts that is in agreement with the main hypothesis, the p-value of progabide treatment is 0.684, indicating it is not significant at $\alpha = 0.05$, given the effect of all other covariates in the model. The same can be said for the treatment by period effect, which has a p-value of 0.646. The lack of significance is reflected in the 95% confidence intervals of the coefficients: $(-0.437, 0.286)$ for treatment and $(-0.166, 0.103)$ for treatment by time. Both contain the zero value, indicating these treatment effects could scale the expected seizure counts by a factor greater than or less than one. Thus, we cannot conclusively state that progabide has an effect on biweekly seizure counts.

The baseline (p-value $< 2 \times 10^{-16}$) and age (p-value $= 0.045$) have significant effects on expected counts. This is expected for the baseline covariate as viaul scans of the clinical trial data show individuals with high biweekly seizure counts tend to have high baseline seizure values. The connection between age and biweekly seizure counts is less clear.

Finally, for our model, the scale parameter estimate of 5.11 indicates that our assumption of the count data following a poisson distribution is incorrect. The observed data has variance exceeding the mean for each patient, indicating that overdispersion occurs in the response variable.

Report Task 2

In part 2, the question regarding results seen in part 1 is: "Does the GEE framework used in model 1 provide accurate coefficient estimates given the small sample size of the clinical trial dataset?"

Experiment Outline

We simulated large data sets with characteristics similar to the trial data and we fit the GEE from part 1 on the simulated data. This allows the GEE framework to provide coefficient estimates that are "good" approximations of asymptotic results obtained from a large patient population. The more often model 1 coefficients fall within the 95% confidence intervals (CI) of their respective coefficients estimated in the simulated data, the more robust the GEE framework is to small samples. Additionally, we ran a quasi-Wald test to study if the set of coefficients in model 1 are seen in 95% confidence regions of each simulation, to investigate frequency of these coefficients occurring together in simulated data.

The data is structured as such: each simulation contains X patients with 4 observations per patient. A patient has 3 fixed variables: Age, Baseline and Treatment and 2 changing variables: Period and Biweekly Seizure Counts (or "counts"). Age, Baseline and Treatment are the same for every observation belonging to one patient, while Time will take values of 1 to 4 for a patient. Counts will take values from a patient-specific dependent count data generating process, specified below.

Data-generating Mechanism & Parameter Justification

To generate Age and Baseline data, we matched the histogram of the empirical distributions to the shape of a Beta and Lognormal distribution respectively. This was done visually for age data and with moment-matching for the baseline data. The estimated distribution parameters were used to simulate new background information for each patient. For treatment status, half of the patients were randomly selected and assigned to the treatment group, with treatment set as one, and the placebo group had treatment set as zero. The period variable for each patient takes values one to four. Using these generated variables, we then create a value:

$$log(\mu_{ij}) = \beta_0 + \beta_1(\text{Treatment}_i) + \beta_2(\text{Time}_j) + \beta_3(\text{Treatment}\cdot\text{Time}_j) + \beta_4(\text{Age}_i) + \beta_5(\text{Baseline}_i) \quad (1)$$

for $i \in \{1, ..., Y\}$, $j \in \{1, 2, 3, 4\}$, where β represents the coefficients estimated for the respective covariates from model 1 in part 1 and μ_{ij} represents the expected biweekly seizure count for patient i at time j. Model 1 coefficients are used as a "benchmark", where if any simulation produces coefficient estimates that are different from those seen in model 1, these differences should not be attributed to covariate effect mis-specification.

To simulate biweekly seizure counts, we generate 4Y random standard normal variables, Z_1, and Y random standard normal variables with each element repeated 4 times, Z_2. To generate dependent standard normals, Z_1 and Z_2 are summed element-wise: $Z = Z1 + \alpha Z2$ to give a vector Z, where α indicates the level of overall within-patient dependence obtained from model 1's estimated correlation parameter. We pass Z into the cumulative density function of a standard normal and obtain the corresponding quantiles Q_z. Q_z is fit into the inverse cumulative mass function of a negative binomial distribution to generate dependent count values. The parameters of the negative binomial distribution has parameters $p = \frac{\mu_{ij}}{\sigma^2_{ij}}$

and $n = \frac{\mu_{ij}^2}{\sigma_{ij}^2 - \mu_{ij}}$, where μ_{ij} is from equation (1) above and $\sigma_{ij}^2 = \mu_{ij} + \frac{\mu_{ij}^2}{\phi}$, ϕ being the empirical dispersion estimate from the clinical trial data.

This method allows us to simulate data similar to the original data by: 1) reflecting the patient's expected counts using their baseline, age, treatment status, treatment duration and model 1 coefficients (μ_{ij}), 2) modelling the overall dependence of within-patient counts using model 1's correlation estimate of within cluster correlation (α), and 3) incorporating the estimated empirical dispersion of the counts in the trial data by using ϕ and a negative binomial distribution, which can generate count data with dispersion greater than 1.

Simulation-specific numbers

We compared the variance of estimated coefficients over 100 or 1000 simulations of 250 patients. In general, 1000 simulations provide a larger variance for coefficient estimates (Figure 1.4). More simulations allow us to capture more variation in the estimates by sampling more from the "theoretical" patient population. Then we tested the number of patients per simulation by comparing variances of each coefficient estimate using 1000 simulations with 100, 250 and 500 patients per simulation (Figure 1.5). We observe that variance of coefficient estimates generally decrease as patient number increases.

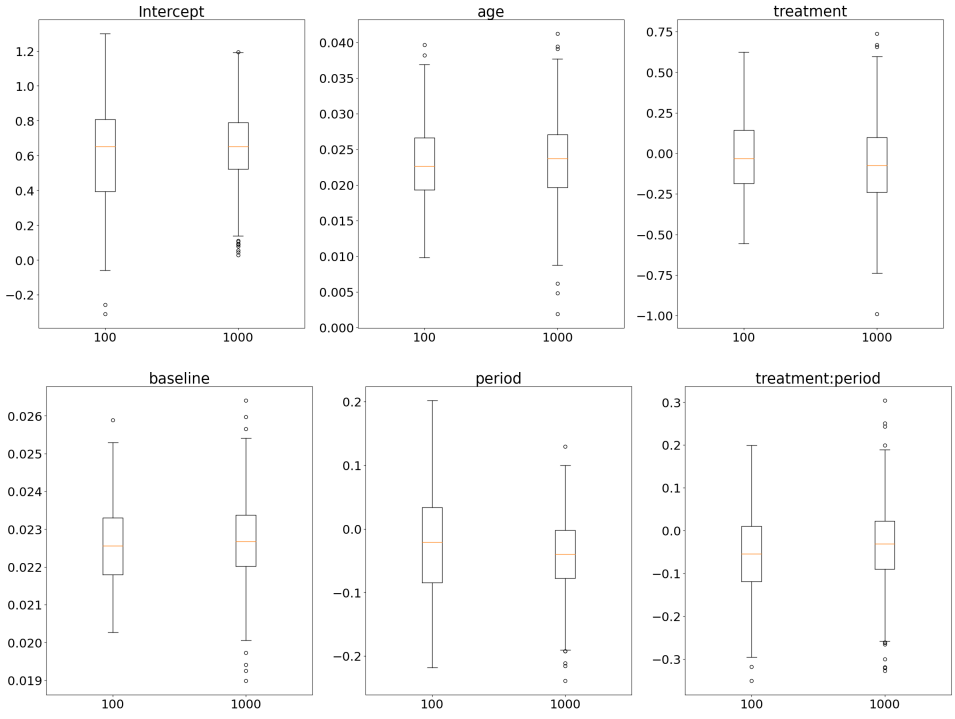

Figure 1.4. Boxplots of coefficient estimates obtained from 100 vs 1000 simulations (of 250 patients)

We selected 500 patients per simulation and 1000 simulations to run for our experiment. Larger values would incur longer computation times, with marginal improvement to estimations.

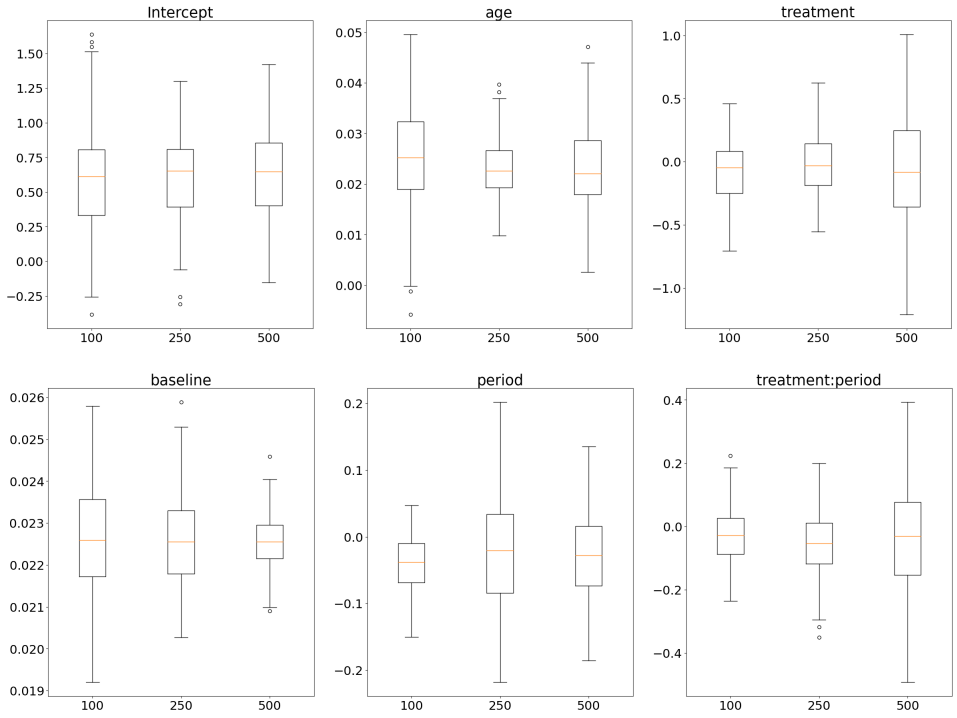

Figure 1.5. Boxplot of coefficient estimates obtained from 100 vs 250 vs 500 patients (over 1000 simulations)

Assuming the GEE framework works well in small samples, we would expect model 1 coefficients to appear with a frequency of 0.95 in the 95% CI constructed for each coefficient in every instance of the simulated data, across the 1000 simulations.

Results & Conclusion

Table 1.2. Table of frequencies of model 1 coefficients occurring in 95% CIs of simulated data coefficients

Number of Simulations	Number of Patients	Intercept	age	treatment	baseline	period	treatment:period
1000	500	0.824	0.888	0.758	0.387	0.841	0.758

From the experiment results, we calculated the frequencies at which model 1 coefficients occur in the constructed 95% CIs for the simulation coefficients. They are presented in Table 1.2. We see all model 1 coefficients occur in their respective simulation 95% CIs less than expected frequency of 0.95. These results indicate the clinical trial data's sample size is too small for the GEE framework. The coefficient estimates produced may not reflect actual estimates obtained from a larger sample of "patients" with similar data structure.

Table 1.3. Table of frequencies of model 1 coefficients occurring in 95% confidence regions of simulated data coefficients

Number of Simulations	Number of Patients	Percentage (confidence_interval=0.95)
1000	500	0.223

For the quasi-Wald test results, out of 1000 simulations, the set of model 1 coefficients occur in only 233 of the 95% confidence regions constructed from estimated coefficients from the simulated data (Table 1.3). This provides more evidence that model estimates obtained from the clinical data may not reflect estimates seen in larger samples.

References

1. Hauser, W. A. Seizure disorders: the changes with age. *Epilepsia* **33,** 6–14 (1992).
2. Sperber, E., Veliskova, J., Germano, I., Friedman, L. K. & Moshe, S. Age-dependent vulnerability to seizures. *Advances in neurology* **79,** 161–169 (1999).