**BIOS6643 Homework 6, Due Nov. 11 at 5pm Trent Hawkins**

1. ***Fit the data using GEE, using the negative binomial distribution and AR(1) working covariance structure. In one or two sentences describe strengths and limitations of this modeling approach in terms of accounting for correlated data.***

Perhaps the largest weakness of GEEs is the inability to specify random effects. The benefit is that it is a computationally easy method to work with. Something worth noting that could be viewed as a benefit or disadvantage is that this model has a population-level interpretation.

1. ***If we had alternatively used a Poisson distribution in (1) and set the dispersion parameter to 1 (i.e., not include a dispersion parameter), how would results have been impacted? (You could also fit the Poisson but include a dispersion parameter, for comparison.)***

If we had used a Poisson distribution, we would not have properly accounted for the overdispersion present in this dataset (see HW 5). Standard Errors would be underestimated, and confidence intervals would be misleadingly narrow.

1. ***Fit the data using GzLMM with quadrature methods with the negative binomial distribution and a random intercept for families. In one or two sentences describe strengths and limitations of this modeling approach.***

The primary benefit of quadrature methods is that one may obtain a true likelihood interpretation. The most major limitation is the inability to apply a more complex structure for the ‘R’ covariance matrix (i.e. repeated effects).

1. ***Fit the data using the GzLMM with pseudo-likelihood methods with the negative binomial distribution and AR(1) covariance structure for repeated measures, plus a random intercept for families. In one or two sentences describe strengths and limitations of this modeling approach.***

The primary benefit of the pseudo-likelihood methods is the ability to specify both random and repeated effects. The largest drawback of this method is that it does not have a true likelihood interpretation, and therefore, it can be difficult to compare to other models.

1. ***Create a table to compare results for the fixed effects in the model (see below). Make sure models use the same reference group. NOTE: you will not have entries for covariance parameters for some modeling approaches (e.g., GEE does not have subject or residual variance terms).***

*All results reported as log (OR)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Parameter** | **GEE** | | **GzLMM (Quad)** | | **GzLMM (PL)** | |
| **Fixed effects** | Estimate | SE | Estimate | SE | Estimate | SE |
| *Intercept* | 1.4531 | 0.1640 | 1.0874 | 0.1416 | 1.2059 | 0.1397 |
| *Week (Control)* | 0.0136 | 0.0087 | 0.01541 | 0.005914 | 0.01154 | 0.007304 |
| *Condition (Experimental Group)* | 0.7349 | 0.1276 | 1.0728 | 0.1529 | 0.9683 | 0.1451 |
| *Week X Condition (Experimental)* | -0.0116 | 0.0093 | -0.02415 | 0.006496 | -0.01760 | 0.008076 |
| *Sex (Female)* | -0.2756 | 0.0663 | -0.2592 | 0.02195 | -0.2563 | 0.03145 |
| *Weight (100 lbs)* | -0.0706 | 0.0681 | -0.07051 | 0.02148 | -0.09830 | 0.03039 |
| *Week (Experimental Group)* | 0.0020 | 0.0035 | -0.00874 | 0.002693 | -0.00606 | 0.003446 |
| **Covariance parameters** | | | | | | |
| *Variance – subject* | NA | | 0.3280 | | 0.2603 | |
| *Variance – residual* | NA | | NA | | 0.5489 | |
| *AR(1)* | 0.756 | | NA | | 0.5034 | |
| *Scale / dispersion* | 0.9583 | | 0.1848 | | 0.3545 | |

1. ***Write a short paragraph summarizing the results of comparing conditions over time.***

All the models make different predictions for the effects that conditions and time have on the outcome. We will interpret the results from the pseudo-likelihood model as it is the only model that can account for both random and repeated effects. All the estimates and confidence intervals have been exponentiated to achieve a relative interpretation.

Holding time constant, the average subject in the experimental group at baseline eats 3 (95% CI: 1.98, 3.50; p < 0.001) times the bowls of cereal as the average subject in the control group. The effect of time does vary, however, by group. For each additional week, the control group eats 1.001 (95% CI: 0.9972, 1.0262; p = 0.1141) times the bowls of cereal (<1% more than the previous week) and the experimental group eats 0.994 (95% CI: 0.9873, 1.001; p = 0.078) times the bowls of cereal (< 1% less than the previous week). Both confidence intervals overlap 1, so the effect of time cannot be considered significant for either group.

1. ***Write a short paragraph comparing the model fits, e.g. differences between models.***

Since these models are not all fit using the same method for likelihood estimation, they will need to be compared in a more qualitative manner. Overall, the models seem to agree on both the magnitude and directions of the fixed effects. Where they seem to differ, most is in the estimates of standard errors. We would expect this outcome since they all estimate likelihood differently, and one method only estimates pseudo-likelihood. Also, these methods can account for different effects. The GEE accounts for only repeated effects. Quadrature accounts only for a random intercept for family. The pseudo-likelihood model is the most flexible and allows us to account for both random and repeated effects. For that reason, I would be most comfortable with the standard error estimates from the pseudo-likelihoods model. Its only disadvantage is that it does not have a true likelihood interpretation, making it difficult to compare to the other models.

1. ***Going back to the model, do you think the straight-line assumption for the time effect appropriate? How can you test for this? (I am open to suggestion, whether you describe a more quantitative approach or something more visual.)***

Chart, line chart

Description automatically generated

Above is a plot of family averages by week for 10 randomly sampled families. Each family’s trajectory is obviously non-linear. This model should adjust for higher-order terms for time. We could test for these trends by specifying time as a class variable and fitting the appropriate contrast statements to find the appropriate term order to include. As we add more terms for time, however, we should be sure that we are adequately powered to detect up-to the non-linear term that we are including.