STA 224 Final Project:
Beta-Carotene and Skin Cancer Prevention
-Application of Longitudinal Analysis

Heqiao Ruan
email:hruan@ucdavis.edu
Instructor: Prof Xiaodong Li

January 27, 2019

1 Abstract

In this project, we apply various techniques in longitudinal data analysis:Generalized Linear Mixed model(GLMM),Generalized Estimating Equation(GEE) to explore the effect of beta carotene on preventing the appearance of the skin cancer as well as corresponding diagnostics to identify the effect of beta-carotene on skin cancer restricted in the first center in the whole data. For these model fitting, we also perform the corresponding model diagnostics to identify the outliers(241th object) which is validated by fitting trajectory visualization. Then we compare the model performance after removing the outliers. Finally we conclude that in this case, Generalized Linear Mixed Model(GLMM) seems to be better than the Generalized Estimating Equation.

2 Introduction

2.1 Background

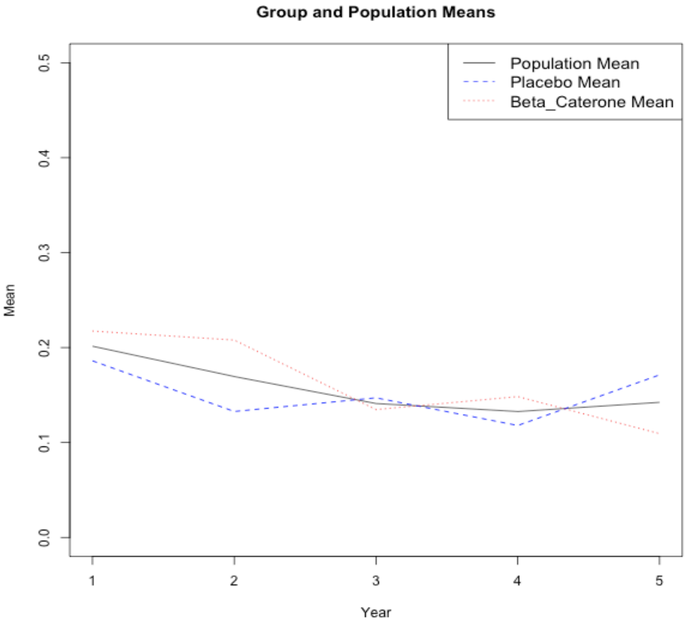
Skin cancer is among one of the most dangerous cancer in the modern society and many research has focused on identifying some specific chemical materials to help prevent the appearance of skin cancer. In 1990, Greenberg([1]) et.al conducted a clinical trial on randomly assigned 1805 people with previous history of relative cancer(denote as high risk subjects) to placebo and beta-catotene group and observe the number of patients' new skin cancer since the previous observation each year in a duration of five years. The ultimate goal of this clinical trial can be interpreted as that beta-carotene has somewhat degree of positive effect to prevent non-melanoma skin cancer in high risk subjects. The beta-carotene is an important kind of nutrients and some previous research shows that it can help cure some kinds of cancer or even prevent the cancer. So here we conduct various methods in longitudinal data analysis to explore the effect of beta-carotene on the skin cancer prevention.

2.2 Dataset and Exploratory Analysis

Here the original dataset we get from the Internet consists of 7081 measurements,1683 subjects with 6 covariates:**Center,Age,Skin,Gender,Exposure,Treatment** where the response variable is in the count form which denotes the number of new skin cancer occurs since the last observation.

One thing we should point out is that it seems impossible for us to accurately identify the difference of effect among different centers which means the covariate **Center** should be excluded from our downstream analysis.So we may need to remove some of the samples from the whole dataset with 4 centers. We solve this issue by restricting our analysis into center 1 which has 422 objects and 1883 measurements in total after pre-processing.

Then we draw the plot of the mean of response among different treatments(group means for placebo group and beta-caterone group) and the whole population mean in one plot after all of the missing data points:



By only observing the plot, we can't tell whether the beta-carotene treatment has significant effect on prevention of skin cancer so we need some quantitative tools to fit the data and perform corresponding diagnostics.

3 GLMM

Here the design is random which means individuals are assigned randomly to two treatment groups and the response variable Y_{ij} here denote the number of new skin cancers for j th year's observation of the ith subject.

3.1 Model selection

The response variable denotes the number of the skin cancers observed from the previous exam and it can be treated as the count variable so the reasonable link function here is chosen as log link(Poisson). Here in fitting the generalized linear mixed model, we may need to start from the full model and prune it by the local wald test(or alternatively, likelihood ratio test). Then we first fit the full model with random slope and random intercept (including most possible covariates),here we use the random intercept and slope b_i which follows a bivariate normal distribution(covariance matrix 2 by 2)

$$log(E[Y_{ij}|b_i]) = \beta_0+\beta_1 trt_i+\beta_2 year_{ij}+\beta_3 age_i+\beta_4 skin_i+\beta_5 gender_i+\beta_6 exposure_i + \beta_7 trt_i * year_{ij} + b_{0i} + b_{1i} * year_{ij} \ \{1\}$$

Then from the coefficient table we can see that effect of $trt_i, year, skin, trt_i * year_{ij}$ are not significant(z test p value much larger than 0.05). What's more,we also conduct a likelihood ratio test which null hypothesis is $H_0 : \beta_1 = \beta_2 = \beta_4 = \beta_7 = 0$ and then the p value for this test is 0.956 which means we can't reject H_0 so then we get the second model:

$$log(E[Y_{ij}|b_i]) = \beta_0 + \beta_1 age_i + \beta_2 gender_i + \beta_3 exposure_i + b_{0i} + b_{1i} * year_{ij} \ \{2\}$$

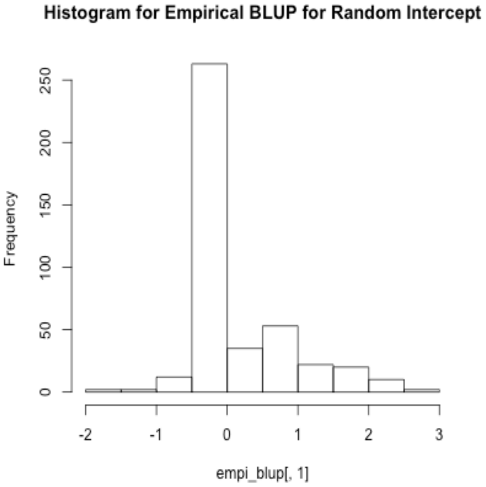
After fitting this model,we observe that all coefficient is significant. Then we need to test the significance of random slope, so we conduct another likelihood ratio test where the reduced model in which only have random intercept but no random slope so here the null hypothesis is $H_0 : b_{1i} = 0$. Then the p value for this likelihood ratio test is 0.000552 which means we reject H_0. So we use the model {2} as our final model for GLMM. Then the model estimation including both the fixed effect and random effect for the generalized linear mixed model is shown as below:

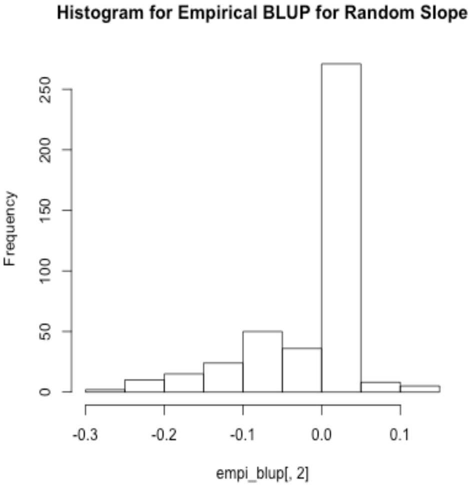
	Estimate	Standard.error	p value
Intercept	-4.758	0.655	3.66e-13
age	0.02126	0.00969	0.028
gender	0.6533	0.1882	0.000518
exposure	0.1714	0.0213	7.46e-16
$var(b_{0i})$	1.9657	_	_
$var(b_{1i})$	0.0225	_	_
$cov(b_{0i}, b_{1i})$	-0.210397	_	_

In terms of model interpretation, we can see that for GLMM, older patients are more likely to develop the cancer no matter what kind of treatment they receive. Similarly, we can see that the log of the expected number of new skin cancers for subjects with more previous skin cancers are larger. What's more, we observe that the male subject has a significant larger log of the expected number of skin cancer than the women subject by the degree of 0.6533. For the random effect, approximately 95% of subjects have changes in expected number of new skin cancers varies from [-0.294,0.294] and b_{0i} here indicates that the subject-wise variability is significant for their log expected number of new skin cancer.

3.2 Diagnostics for GLMM

For diagnostic part of GLMM, we first use the χ^2 test by the sum of square of pearson residual to check the model adequancy(i.e. whether overdispersion exist).Then the test statistic here is $\sum_{i,j} r^2_{P,ij} = \sum_{i,j} \frac{(Y_{ij}-\hat{\mu_{ij}})^2}{\hat{\mu_{ij}}}$. Here we know that under the null hypothesis(model is adequate), it follows χ^2_{n-p}. For the degree of freedom, it is the total number of parameters in the model which include both the random effect and fixed effect. Then the p value here is 1 which means we don't have enough evidence to reject the null hypothesis. So we conclude that the model is adequate.
What's more, we plot the empirical BLUP to see whether there are outliers:





We can see that there may be an outlier and after investigation, it's the 241th object. The following analysis in GEE validate this observation.

4 GEE

For fitting the generalized estimating equation(GEE), we also use the poisson family(count response).Similarly we will start from a full model which include all possible covariates and their interactions and we assume the cubic time trend.What's more, WLOG, we assume our correlation pattern as unstructured. So the first model here is

$$log(E[Y_{ij}]) = \beta_0+\beta_1 year_{ij}+\beta_2 age_i+\beta_3 skin_i+\beta_4 gender_i+\beta_5 exposure_i + \beta_6 trt_i * year_{ij} + \beta_7 trt_i * age_i + \beta_8 trt_i * skin_i + \beta_9 trt_i * gender_i + \beta_{10} year_{ij}^2 + \beta_{11} year_{ij}^3 + \beta_{12} trt_i + \beta_{13} trt_i * exposure_i \; \{3\}$$

Then from the coefficient table we can see that only $trt, exposure, I(trt * age), I(trt * gender)$ are significant. So here we conduct the likelihood ratio test for the null hypothesis $H_0 : \beta_{rest} = 0$ where β_{rest} denote all of the rest variables apart from the four variables. So here comes our second model:

$$log(E[Y_{ij}]) = \beta_0 + \beta_1 trt_i + \beta_2 exposure_i + \beta_3 trt_i * age_i + \beta_4 trt_i * gender_i \; \{4\}$$

Then the p value for the LR test is 0.34 which favors the reduced model.
After reaching a reasonable model, we would like to try different correlation patterns: AR(1), exchangeable, unstructured and independent.From [3](Lab Note of UNC ECOL562) we know that, to select the best pattern, we can just choose the one with the smallest deviation between the naive correlation estimation and the robust correlation estimation.(Alternative methods includes QIC or empirically observe the data or domain knowledge) To evaluate the difference, we may use the l_1 norm of the matrix which sums the absolute value of all elements in the working correlation matrix. Here a better fit means that our model-based estimator is close to our sandwich estimator achieved by the GEE procedure:difference of Robust(sandwich) s.e and naive s.e(Sandwich estimator most approximate the NAIVE estimator). Then the difference of the two estimators for different correlation structures are shown as below.

Unstructured	AR(1)	Exchangeable	Independent
0.242	0.662	0.426	0.838

So here we may choose the unstructured covariance structure.Then model summary is shown as below:

Variable	Estimate	Standard.error	p value
Intercept	-2.4838	0.1324	$< 2e-16$
trt	-3.9569	1.0547	0.000176
exposure	0.13668	0.01197	$< 2e-16$
I(trt*age)	0.0521	0.0152	0.000589
I(trt*gender)	1.0419	0.3044	0.000620
ϕ	1.27	0.278	-

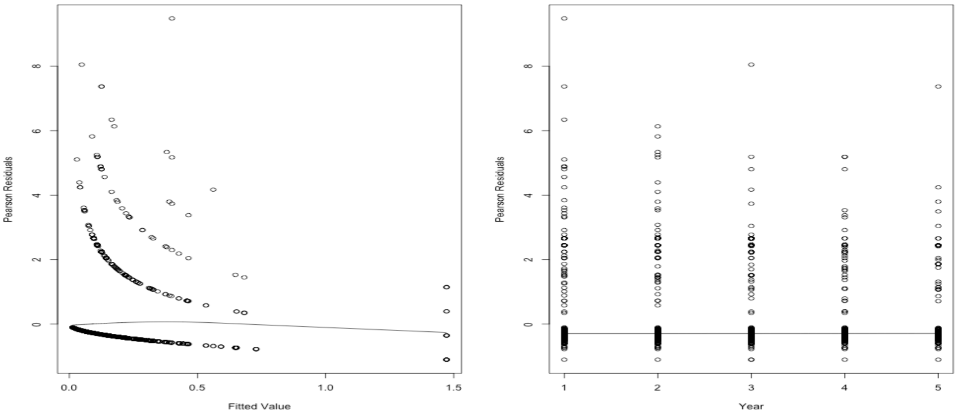
The correlation structure here $Corr(Y_{ij}, Y_{ij})$ is estimated as a unstructured matrix and variance of sample is defined as $Y_{ij} = \phi E[Y_{ij}]$.Then the estimated correlation structure is shown as below:

$$\begin{bmatrix} 1.00000 & 0.32462 & 0.15279 & 0.29707 & 0.18767 \\ 0.32462 & 1.00000 & 0.09347 & 0.13870 & -0.00246 \\ 0.15279 & 0.09347 & 1.00000 & -0.00953 & 0.13381 \\ 0.29707 & 0.13870 & -0.00953 & 1.00000 & -0.04172 \\ 0.18767 & -0.00246 & 0.13381 & -0.04172 & 1.00000 \end{bmatrix}$$

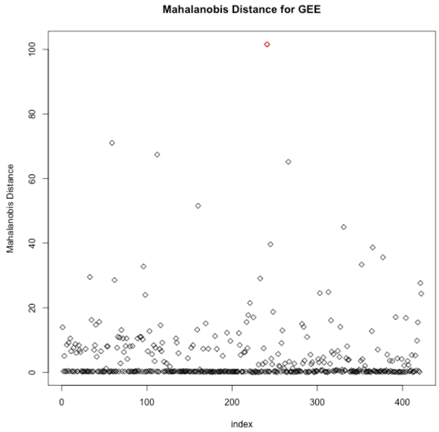
So here we can see that the beta-carotene treatment has significant effect on reducing the expected number of new skin cancers detected while the high exposure will contribute to the relapse of skin cancer. What's more, we can see that beta-carotene works better to reduce the risk of developing new cancers in female and younger patients. Here the $\hat{\phi}$ is only slightly larger than 1 means that there are no significant degree of overdispersion.

4.1 Diagnostics for GEE

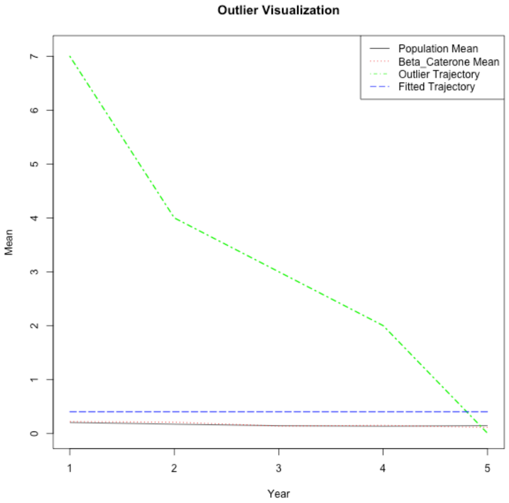
Here from CH13,[2], we use studentized pearson residuals $e_{ij} = \frac{Y_{ij}-g^{-1}(X_{ij}^T\hat{\beta})}{\sqrt{\phi v(\hat{\mu_{ij}})}}$ and the mahalanobis distance between the observed samples and fited values $d_i = r_i^T\hat{V^{-1}}r_i$ (where $\hat{V^{-1}}$ is the estimated covariance structure) to perform model diagnostics. Then we draw the residual versus fitted value and residual versus year plot. From the plots below we can see that there's no systematic trend for the fitted curves so it indicates that there's no lack of fit.



Then we plot the mahalanobis distance(it just remove the heterogeneity) to do the outlier detection. We treat the observation with the largest mahalanobis distance as red:



We observe that the 241th object has the largest mahalanobis distance so it may probably be the outlier case. Then we plot the trajectory for this object comparing with the global and its treatment groups' mean and the plot is shown as below:



We can see that obviously, the 241th observation is an outlier. Then we remove this object and fit the GEE again. Then the final model we choose is similar to that before removing outliers, mainly differ in the covariance structure, here we use the exchangeable covariance structure by applying the similar procedure of structure choosing as above from [3].

Variable	Estimate	Standard.error	p value
Intercept	-2.3893	0.1315	$< 2e - 16$
trt	-3.1604	0.9479	0.00086
exposure	0.1289	0.0117	$< 2e - 16$
I(trt*gender)	0.9087	0.2879	0.00160
I(trt*age)	0.0402	0.0136	0.00317
ϕ	1.13	_	_
ρ	0.104	_	_

The model fitted after removing outliers are similar to that before removing because the sign of variables are all the same and the magnitude of coefficients and the estimated scale parameter $\hat{\phi}$ are also very similar.

5 Model Comparison

Here for both the GLMM and GEE models, we do some visualizations which can also be denoted as trajectory display, we randomly select 16 objects from the dataset and draw the fitted lines by both the GLMM and GEE model and compare them(See Appendix), we can see that the model by GLMM is slightly more efficient than the GEE because the fitted trajectory is more like the observation points(for example the 3th, the 7th and the 11th random selected observation. Indeed it is reasonable because GLMM also include the random effect for each subject which is not explored by GEE. **What's more, from CH13[2] we know that if the missing pattern is MAR(missing at random) not MCAR(missing completely at random), we can see that in this case the GEE model may not be as efficient as in usual. But in this case, GLMM in R indeed use the maximum likelihood method which is not influenced by the missing pattern.**
So we conclude here that Generalized Linear Mixed Model fits the dataset better than the Generalized Estimating Equation.

6 Discussion

In this report for analyzing the skin cancer prevention study data, we compare the two methods very popular in the longitudinal data analysis: GLMM and GEE. We identify that the 241 th observation as the outlier case and validate it by visualizing the trajectory. Then not surprisingly, after conducting these two methods, we get pretty similar results. However, from the trajectory visualization, we can see that GLMM also include the subject-wise effect which borrows more information from the data points than the GEE which can be interpreted as marginal model. For the effect of beta-carotene, we can see that it can significantly reduce the risk of developing new skin cancer for patients. What's more, we can see that patients with more previous cancers are significantly more likely to develop new cancer.

References

[1] Greenberg,E.R.,Baron,J.A.,Stukel,T.A.,Stevens,M.M.,Mandel,J.S.,Spencer,S.K.,Elias,P.M. ,Lowe,N.,Nierenberg,D.W.,Bayrd,G.,Vance,J.C.,Freeman,D.H.,Clendenning,W.E.,Kwan,T. and the Skin Cancer Group(1990) *A clinical trial of beta carotene to prevent basal-cell and squamous-cell cancers of the skin.* New England Journal of Medicine,323,789-795.

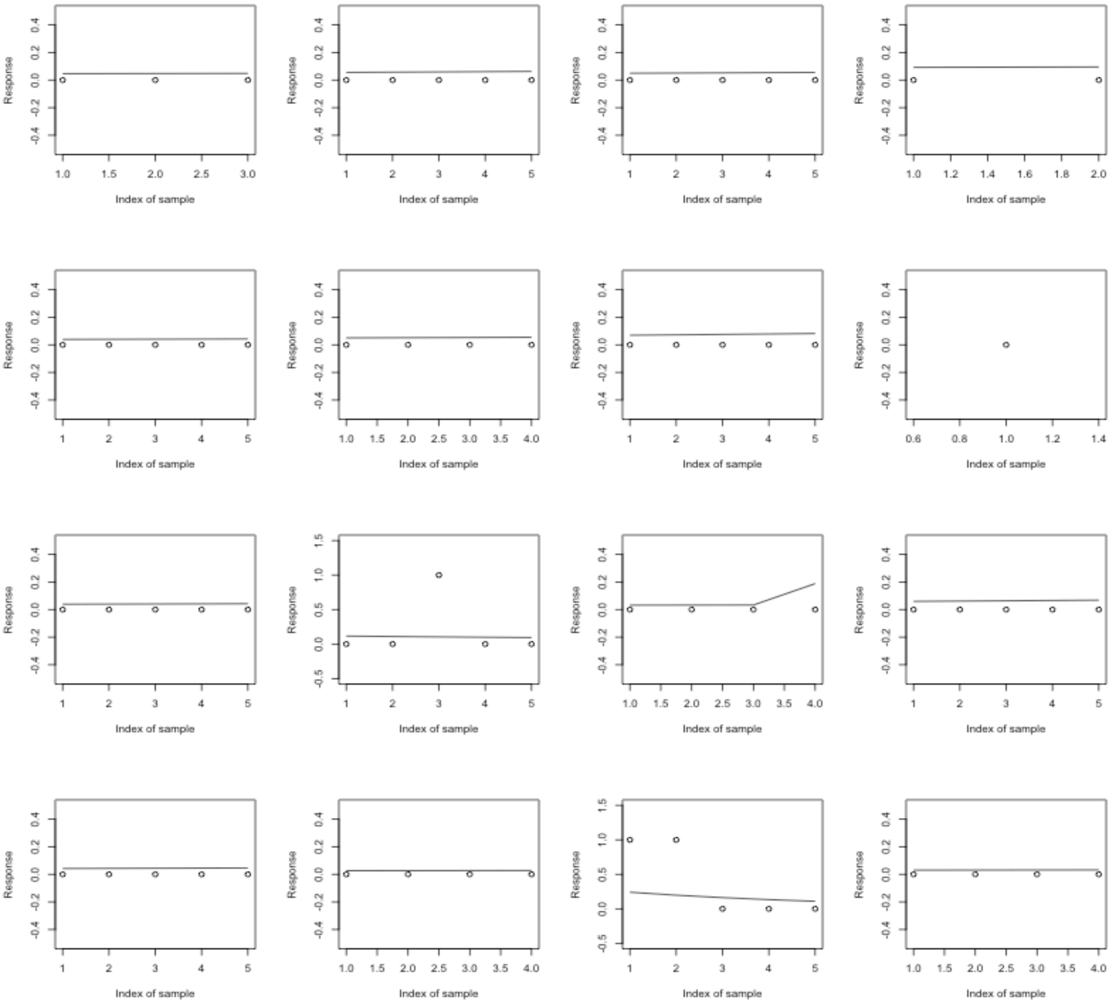
[2] Garrett M.Fitzmaurice,Nan M.Laird and James H.Ware. 2011. *Applied Longitudinal Analysis* John Wiley,Sons,Inc.,Hoboken,New Jersey.

[3] UNC ECOL 562 *Lecture 13* $https : //www.unc.edu/courses/2010spring/ecol/562/001/docs/lectures/lecture13.htm\#choosing$

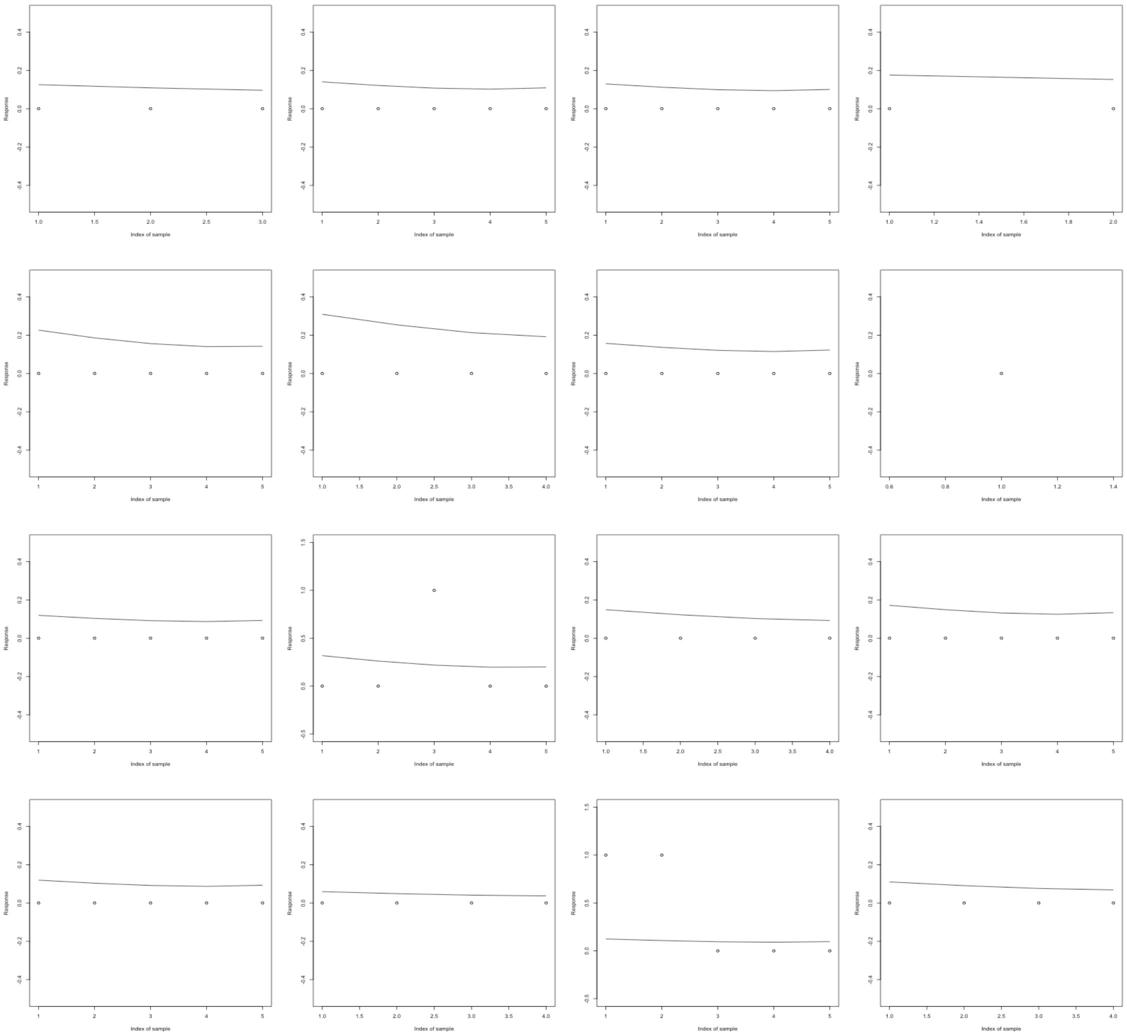
7 Appendix

7.1 Plots

Trajectory Visualization for GLMM models:



Trajectory Visualization for GEE models:


Response
Index of sample