PharmaSUG China 2023 - Paper AP-172

SAS vs R: Comparison and Implementation on Mixed Model Repeated Measures (MMRM) and Negative Binomial Regression

Xinya Wang, Xinying Chen, AstraZeneca

ABSTRACT

In recent years, there has been a growing interest in utilizing R as an alternative to SAS for statistical analysis for clinical studies trials. R, as an open-source programming language, offers a wide range of statistical techniques, data manipulation capabilities, and visualization possibilities. Its extensive library enables researchers to access and implement cutting-edge methodologies easily.

This paper presents a comprehensive comparison of statistical modeling of used in clinical trials, which are Mixed-Effect Models for Repeated Measures (MMRM) and Negative Binomial regression, by two popular software: SAS and R, demonstrates the transition from SAS to R programming, and highlights the challenges and practical considerations of using R for statistical analysis.

The differences in coding syntax and results interpretation will be emphasized based on a dummied clinical trial example data, aiming to provide researchers, statisticians, and programmers with valuable insights into using R for statistical modeling.

INTRODUCTION

MMRM and Negative Binomial regression are commonly used statistical models in clinical trials.

For MMRM, PROC MIXED procedure in SAS and the mmrm package in R is commonly used to fit MMRM models.

For Negative Binomial regression, PROC GENMOD is the procedure used in SAS for analysis while package MASS and emmeans are commonly used in R for similar analysis purpose.

In this paper, we will discuss the main procedures and packages used to perform modeling analysis in SAS and R, summarize analysis results and add detailed interpretations, then provide an overall comparison between SAS with R for both of above statistical modeling.

MMRM

MMRM stands for Mixed-Effects Model Repeated Measures, which is a statistical method used to analyze data in longitudinal or repeated measures clinical trials

OVERVIEW OF MMRM

MMRM typically does not specify patient-level random effects, but instead models the correlation within the repeated measures over time by assuming that the residual errors are correlated. To reduce the chances of model misspecification, the residual errors are often assumed to be from a multivariate normal distribution with an unstructured covariance matrix. This allows for flexibility in the form of the correlation matrix of the repeated measures. (Geert Molenberghs, 2004)

The fixed effects in an MMRM include effects of time (as a categorical or factor variable), randomized treatment group, and their interaction. This implies a saturated model for the mean, where there is a separate mean parameter for each time point in each treatment group. Baseline covariates can also be included in the model as simple main effects or with an interaction with time to allow for potential variation in the association between the baseline variable(s) and outcome over time. (Geert Molenberghs, 2004)

One of the key benefits of using an MMRM is that it can handle missing data in a flexible way, because the model assumes that missing data is at random, it can use all available data to estimate the model parameters. This is particularly useful in clinical trials where missing data is common due to patient drop-out or other reasons. (Geert Molenberghs, 2004)

ANALYSIS

In this section, we will display the example data and MMRM modeling information, provide main procedures and outputs interpretation base on SAS, main R packages and outputs, and comparison between R and SAS.

DATA AND ANALYSIS PURPOSE

Here we dummied a study with a Regional, Multicenter, Randomized, Double-Blind, Placebo Controlled, Parallel Group, to Evaluate the Efficacy of Drug X in Adults with Asthma. In the study, 400 subjects are randomized into Drug X and placebo, with visits at week 1 to week 9 to collect the results of total score of Standardized Asthma Quality of Life Questionnaire for 12 Years and Older(AQLQ(S)+12).

Change from baseline for AQLQ(S)+12 as the study endpoints in the Drug X group will be compared to that seen in the placebo group using MMRM modeling. In this model, inference will be based on the restricted maximum likelihood (REML) estimation. The response variable in the model will be the change from baseline at each scheduled post baseline visits. Treatment, visit, region, and treatment by visit interaction will be included as factors in this model. Baseline of the corresponding endpoint will also be included in the model as a continuous l covariate. Unstructured covariance will be assumed to model the relationship between pairs of response variables taken at different visits on the same subject. The Kenward-Roger approximation to estimating the degrees of freedom will be used for tests of fixed effects derived from this model above.

SAS CODE, OUTPUTS, AND INTERPRETATION

The main procedure used in SAS to fit MMRM model will be based on proc mixed.

PROC MIXED PROCEDURE

The MIXED procedure is a powerful tool for fitting mixed linear models to data and making statistical inferences. This procedure is more general than GLM in the sense that it gives a user the flexibility of modeling not only the means of your data (as in the standard linear model) but their variances and covariances as well. (Center, 2019)

PROC MIXED provides a variety of covariance structures to generalized to include one set of clusters nested within another, and also longitudinal studies, where repeated measurements are taken over time.

The basic features for MIXED procedure including:

- Different Covariance structures, like variance components, compound symmetry, unstructured, AR(1), Toeplitz, spatial, general linear, and factor analytic.
- GLM-type grammar using MODEL, RANDOM, and REPEATED statements for model specification and CONTRAST, ESTIMATE, and LSMEANS statements for inferences.
- Appropriate standard errors and t/F tests for estimable linear combinations of fixed and random effects.
- Subject and group effects for blocking and heterogeneity.
- REML and ML estimation methods with Newton-Raphson algorithm.
- Handling of unbalanced data.
- Creation of SAS data set corresponding to any tables. (Center, 2019)

SAS CODES

To fit linear modeling for repeated measures to estimate the change from baseline of AQLQ(S)+12, the SAS codes shall be as follows:

```
proc mixed data=adqs method=reml covtest;
   class usubjid avisit trt01p(REF='Placebo') region1 ;
   model chg = trt01p avisit base region1 trt01p*avisit / ddfm=KR;
   repeated avisit / subject=usubjid type=un;
   lsmeans trt01p * avisit / cl pdiff alpha = 0.05  obsmargins;
```

```
quit;
```

- **proc mixed**: is being used to perform a linear mixed effects analysis.
- **method=reml**: specifies to call residual (restricted) maximum likelihood to provides less biased estimates of the variance components of the model.
- **covtest** option: requests asymptotic standard errors and Wald-Z tests for covariance parameters. This option has SAS show hypothesis tests for the variance and covariance parts of the model in the output.
- **class usubjid avisitn trt01p region1**: specifies the categorical variables in the model.
- **model chg = trt01p avisit base region1 trt01p*avisit:** first lists the response variable chg. The fixed effects are then listed after the equal sign, including trt01p, avisit, base, region1, and their interaction terms.
- **ddfm=KR**: specifies degrees of freedom method using Kenward-Roger.
- **repeated avisit / subject=usubjid**: specifies the repeated measures structure of the data, with "avisit" as the repeated measure variable and "usubjid" as the subject variable.
- **type=un**: requests an unstructured block for each SUBJECT= usubjid.
- **lsmeans trt01p * avisit / cl pdiff alpha = 0.05 obsmargins**: computes least squares means for the interaction between trt01p and avisit, with confidence limits and pairwise differences between means, diff option provides differences between each treatment group and the obsmargins option includes observed margins in the output.

SAS OUTPUTS AND RESULTS INTERPRETATION

The results from the above SAS procedures are shown in table 1- table 9.

Table 1 Model information

Data Set	ADQS
Dependent Variable	CHG
Covariance Structure	Unstructured
Subject Effect	USUBJID
Estimation Method	REML
Residual Variance Method	None
Fixed Effects SE Method	Kenward-Roger
Degrees of Freedom Method	Kenward-Roger

As the above table shows, the covariance structure is listed as “Unstructured,” and no residual variance is used with this structure. The default degrees-of-freedom method here is “**Kenward-Roger**.”

Table 2 Covariance information

Covariance Parameters	45
Columns in X	33
Columns in Z	0
Subjects	400
Max Obs per Subject	9

In table 2, the 45 covariance parameters result from the 9X9 unstructured blocks of R. There is no Z matrix for this model, and each of the 400 subjects has a maximum of 9 observations.

Table 3 Iteration History

Iteration	Evaluations	-2 Res Log like	Criterion
0	1	16114.89197	_
1	2	14759.87639	0.00000255
2	1	14759.86369	0.00000000
Convergence criteria met.			

Three iterations are required to find the maximum likelihood estimates. The default relative Hessian criterion has a final value less than 1E–8, indicating the convergence of the Newton-Raphson algorithm and the attainment of an optimum.

Table 4 Estimated R Matrix for the first subject (USUBJID = AB12345-BRA-1-id-129)

Row	Col1	Col2	Col3	Col4	Col5	Col6	Col7	Col8	Col9
1	1.7094	-0.09411	-0.0786	0.1529	0.334	0.02119	-0.8279	-0.2446	1.6972
2	-0.09411	4.2065	0.1936	-0.767	-0.2298	-0.2552	2.4422	-0.9474	-1.1111
3	-0.0786	0.1936	7.6257	-1.3111	-0.6976	0.6102	0.9898	0.02888	-0.7187
4	0.1529	-0.767	-1.3111	12.6079	3.0353	0.9823	1.4496	-2.5108	2.5796
5	0.334	-0.2298	-0.6976	3.0353	18.1433	-0.09849	1.0156	-2.2994	2.4922
6	0.02119	-0.2552	0.6102	0.9823	-0.09849	27.7964	4.2212	0.05593	-7.1592
7	-0.8279	2.4422	0.9898	1.4496	1.0156	4.2212	41.5924	-5.0348	-0.9494
8	-0.2446	-0.9474	0.02888	-2.5108	-2.2994	0.05593	-5.0348	45.6269	0.444
9	1.6972	-1.1111	-0.7187	2.5796	2.4922	-7.1592	-0.9494	0.444	63.8729

In the above table, the 9X9 matrix is the estimated unstructured covariance matrix. It is the estimate of the first block of R, and the other 399 blocks all have the same estimate.

Table 5 Covariance Parameter Estimates

CovParm	Subject	Estimate	StdErr	ZValue	ProbZ		CovParm	Subject	Estimate	StdErr	ZValue	ProbZ
UN(1,1)	USUBJID	1.7094	0.1427	11.98	<.0001		UN(7,1)	USUBJID	-0.8279	0.5377	-1.54	0.1236
UN(2,1)	USUBJID	-0.09411	0.1708	-0.55	0.5817		UN(7,2)	USUBJID	2.4422	0.9039	2.7	0.0069
UN(2,2)	USUBJID	4.2065	0.3416	12.31	<.0001		UN(7,3)	USUBJID	0.9898	1.1236	0.88	0.3784
UN(3,1)	USUBJID	-0.0786	0.2384	-0.33	0.7417		UN(7,4)	USUBJID	1.4496	1.5723	0.92	0.3566
UN(3,2)	USUBJID	0.1936	0.3494	0.55	0.5795		UN(7,5)	USUBJID	1.0156	1.9444	0.52	0.6015
UN(3,3)	USUBJID	7.6257	0.6154	12.39	<.0001		UN(7,6)	USUBJID	4.2212	2.2152	1.91	0.0567
UN(4,1)	USUBJID	0.1529	0.3234	0.47	0.6363		UN(7,7)	USUBJID	41.5924	3.4334	12.11	<.0001
UN(4,2)	USUBJID	-0.767	0.4863	-1.58	0.1148		UN(8,1)	USUBJID	-0.2446	0.6137	-0.4	0.6902
UN(4,3)	USUBJID	-1.3111	0.7055	-1.86	0.0631		UN(8,2)	USUBJID	-0.9474	0.9355	-1.01	0.3112
UN(4,4)	USUBJID	12.6079	1.047	12.04	<.0001		UN(8,3)	USUBJID	0.02888	1.2602	0.02	0.9817
UN(5,1)	USUBJID	0.334	0.4092	0.82	0.4144		UN(8,4)	USUBJID	-2.5108	1.6794	-1.5	0.1349
UN(5,2)	USUBJID	-0.2298	0.6016	-0.38	0.7025		UN(8,5)	USUBJID	-2.2994	1.9468	-1.18	0.2375
UN(5,3)	USUBJID	-0.6976	0.7851	-0.89	0.3743		UN(8,6)	USUBJID	0.05593	2.3144	0.02	0.9807
UN(5,4)	USUBJID	3.0353	1.0061	3.02	0.0026		UN(8,7)	USUBJID	-5.0348	3.2024	-1.57	0.1159
UN(5,5)	USUBJID	18.1433	1.5187	11.95	<.0001		UN(8,8)	USUBJID	45.6269	3.7666	12.11	<.0001
UN(6,1)	USUBJID	0.02119	0.4536	0.05	0.9627		UN(9,1)	USUBJID	1.6972	0.7595	2.23	0.0254

UN(6,2)	USUBJID	-0.2552	0.7078	-0.36	0.7184		UN(9,2)	USUBJID	-1.1111	1.0243	-1.08	0.278
UN(6,3)	USUBJID	0.6102	0.9444	0.65	0.5182		UN(9,3)	USUBJID	-0.7187	1.5158	-0.47	0.6354
UN(6,4)	USUBJID	0.9823	1.2151	0.81	0.4189		UN(9,4)	USUBJID	2.5796	1.8675	1.38	0.1672
UN(6,5)	USUBJID	-0.09849	1.6489	-0.06	0.9524		UN(9,5)	USUBJID	2.4922	2.3429	1.06	0.2875
UN(6,6)	USUBJID	27.7964	2.2571	12.32	<.0001		UN(9,6)	USUBJID	-7.1592	2.9749	-2.41	0.0161
							UN(9,7)	USUBJID	-0.9494	3.5427	-0.27	0.7887
							UN(9,8)	USUBJID	0.444	3.7819	0.12	0.9065
							UN(9,9)	USUBJID	63.8729	5.3781	11.88	<.0001

The “Covariance Parameter Estimates” table lists the 45 estimated covariance parameters in order; note their correspondence to the first block of R displayed in table 4. The parameter estimates are labeled according to their location in the block in the **Covariance Parameter** column, and all these estimates are associated with USUBJID as the subject effect. These standard errors lead to approximate Wald Z statistics, which are compared with the standard normal distribution. Please note the above statistics constitute Wald tests of the covariance parameters are valid only asymptotically.

Table 6 Fit Statistics and null model likelihood ratio test

Description	**Value**	
-2 Res Log Likelihood	14759.9	
AIC (Smaller is Better)	14849.9	
AICC (Smaller is Better)	14851.5	
BIC (Smaller is Better)	15029.5	
Null Model Likelihood Ratio Test		
DF	**ChiSq**	**ProbChiSq**
44	1355.03	<.0001

The Fit Statistics are relative measures of how well the current model fits the data compared to other model parameterizations. If all the model parameterizations fit the data poorly, the fit statistics will not reflect this. They only indicate which of the models provide the better fit.

The null model likelihood ratio test (LRT) in table 6 is highly significant for this model, indicating that the unstructured covariance matrix is preferred to the diagonal matrix of the ordinary least squares null model. The degrees of freedom for this test is 44, which is the difference between 45 and the 1 parameter for the null model’s diagonal matrix.

Table 7 Least Squares Means

Effect	**AVISIT**	**TRT01P**	**Margins**	**Estimate**	**StdErr**	**DF**	**Lower**	**Upper**
AVISIT*TRT01P	WEEK 1 DAY 8	Drug X	ADQS	2.1834	0.1116	292	1.9637	2.4031
AVISIT*TRT01P	WEEK 1 DAY 8	Placebo	ADQS	1.8301	0.1071	292	1.6194	2.0408
AVISIT*TRT01P	WEEK 2 DAY 15	Drug X	ADQS	3.9373	0.1707	310	3.6015	4.2732
AVISIT*TRT01P	WEEK 2 DAY 15	Placebo	ADQS	4.0627	0.161	310	3.746	4.3794
AVISIT*TRT01P	WEEK 3 DAY 22	Drug X	ADQS	5.5607	0.2286	310	5.1109	6.0104
AVISIT*TRT01P	WEEK 3 DAY 22	Placebo	ADQS	5.9681	0.2174	311	5.5403	6.3959
AVISIT*TRT01P	WEEK 4 DAY 29	Drug X	ADQS	8.2179	0.3099	298	7.608	8.8279
AVISIT*TRT01P	WEEK 4 DAY 29	Placebo	ADQS	7.6612	0.2803	297	7.1096	8.2128
AVISIT*TRT01P	WEEK 5 DAY 36	Drug X	ADQS	9.4389	0.3854	290	8.6803	10.1975
AVISIT*TRT01P	WEEK 5 DAY 36	Placebo	ADQS	9.6264	0.3333	290	8.9703	10.2824

AVISIT*TRT01P	WEEK 6 DAY 43	Drug X	ADQS	11.4273	0.4496	308	10.5426	12.312
AVISIT*TRT01P	WEEK 6 DAY 43	Placebo	ADQS	11.6699	0.4074	308	10.8682	12.4716
AVISIT*TRT01P	WEEK 7 DAY 50	Drug X	ADQS	12.1631	0.5504	300	11.0799	13.2463
AVISIT*TRT01P	WEEK 7 DAY 50	Placebo	ADQS	13.6757	0.5126	300	12.6668	14.6845
AVISIT*TRT01P	WEEK 8 DAY 57	Drug X	ADQS	15.369	0.6	297	14.1882	16.5498
AVISIT*TRT01P	WEEK 8 DAY 57	Placebo	ADQS	15.8817	0.523	296	14.8525	16.911
AVISIT*TRT01P	WEEK 9 DAY 64	Drug X	ADQS	16.2995	0.6958	290	14.9301	17.6689
AVISIT*TRT01P	WEEK 9 DAY 64	Placebo	ADQS	17.9912	0.64	290	16.7315	19.2509

The above table 7 displays the estimates of least square means, Standard error, degree of freedom, lower confidence interval, and higher confidence interval for each treatment group and each analysis visit means respectively.

Table 8 Differences of Least Squares Means

Effect	AVISIT	TRT01P	margins	Estimate	StdErr	DF	tValue	Probt	Lower	Upper
AVISIT*TRT01P	WEEK 1 DAY 8	Drug X	ADQS	0.3533	0.155	293	2.28	0.0234	0.04822	0.6583
AVISIT*TRT01P	WEEK 2 DAY 15	Drug X	ADQS	-0.1254	0.2349	311	-0.53	0.5939	-0.5875	0.3368
AVISIT*TRT01P	WEEK 3 DAY 22	Drug X	ADQS	-0.4074	0.3156	311	-1.29	0.1977	-1.0285	0.2136
AVISIT*TRT01P	WEEK 4 DAY 29	Drug X	ADQS	0.5567	0.418	298	1.33	0.1839	-0.2659	1.3794
AVISIT*TRT01P	WEEK 5 DAY 36	Drug X	ADQS	-0.1875	0.5097	290	-0.37	0.7133	-1.1906	0.8156
AVISIT*TRT01P	WEEK 6 DAY 43	Drug X	ADQS	-0.2426	0.6068	308	-0.4	0.6896	-1.4366	0.9515
AVISIT*TRT01P	WEEK 7 DAY 50	Drug X	ADQS	-1.5125	0.7523	300	-2.01	0.0453	-2.9929	-0.03217
AVISIT*TRT01P	WEEK 8 DAY 57	Drug X	ADQS	-0.5127	0.7959	297	-0.64	0.52	-2.0791	1.0537
AVISIT*TRT01P	WEEK 9 DAY 64	Drug X	ADQS	-1.6917	0.9454	290	-1.79	0.0746	-3.5524	0.1691

Table 8 has a statistics estimate (LS means difference), stdErr (standard error), Probt (P value), Lower (lower confidence interval), and Upper (upper confidence interval). The t-test is used to estimate the difference in mean change from baseline in AQLQ(S)+12 total score at each analysis visit (Drug X minus placebo) = 0, and the above P value shows that there is no statistically significant difference in means for change from baseline in AQLQ(S)+12 for Drug X compared to Placebo.

R PACKAGE, CODES, OUTPUTS, AND INTERPRETATION

The main R package used here to fit MMRM will be based on mmrm package and emmeans package.

mmrm PACKAGE

In early 2022, the mmrm package was developed to address the issues faced by R packages such as lme4, nlme, and glmmTMB in generating MMRM outputs, and incorporate important features such as the Satterthwaite adjusted degrees of freedom and various covariance structures including unstructured (UN), homogeneous Toeplitz (TOEP), auto-regressive (AR(1)), heterogeneous auto-regressive (ARH(1)), compound symmetry (CS), and heterogeneous compound symmetry (CSH), etc..

The mmrm package utilizes the marginal linear model without random effects through Template Model Builder (TMB), resulting in efficient and reliable model fitting. TMB enables C++ log-likelihood calculations, making this package 3-5 times faster than previous options.

Additionally, hypothesis testing with Satterthwaite or Kenward-Roger adjustment is available, and least square means estimates can be extracted using the emmeans package. The mmrm package employs multiple optimizers if the default option does not lead to convergence. Further details on the model fitting algorithm and features are available on the Open Pharma Github page and in the documentation on CRAN. (Sabanes Bove D, mmrm: Mixed Models for Repeated Measures. R package version 0.2.2.9027, 2023)

PARAMCD	Effect	AVISIT	TRT01P	_AVISIT	_TRT01P	margins	Estimate	StdErr	DF	tValue	Probt	Alpha	Lower	Upper
AQLQAASC	AVISIT*TRT01P	WEEK 1 DAY 8	Drug X	WEEK 1 DAY 8	Placebo	WORK.ADQS	0.3533	0.1550	293	2.28	0.0234	0.05	0.04822	0.6583
AQLQAASC	AVISIT*TRT01P	WEEK 2 DAY 15	Drug X	WEEK 2 DAY 15	Placebo	WORK.ADQS	-0.1254	0.2349	311	-0.53	0.5939	0.05	-0.5875	0.3368
AQLQAASC	AVISIT*TRT01P	WEEK 3 DAY 22	Drug X	WEEK 3 DAY 22	Placebo	WORK.ADQS	-0.4074	0.3156	311	-1.29	0.1977	0.05	-1.0285	0.2136
AQLQAASC	AVISIT*TRT01P	WEEK 4 DAY 29	Drug X	WEEK 4 DAY 29	Placebo	WORK.ADQS	0.5567	0.4180	298	1.33	0.1839	0.05	-0.2659	1.3794
AQLQAASC	AVISIT*TRT01P	WEEK 5 DAY 36	Drug X	WEEK 5 DAY 36	Placebo	WORK.ADQS	-0.1875	0.5097	290	-0.37	0.7133	0.05	-1.1906	0.8156
AQLQAASC	AVISIT*TRT01P	WEEK 6 DAY 43	Drug X	WEEK 6 DAY 43	Placebo	WORK.ADQS	-0.2426	0.6068	308	-0.40	0.6896	0.05	-1.4366	0.9515
AQLQAASC	AVISIT*TRT01P	WEEK 7 DAY 50	Drug X	WEEK 7 DAY 50	Placebo	WORK.ADQS	-1.5125	0.7523	300	-2.01	0.0453	0.05	-2.9929	-0.03217
AQLQAASC	AVISIT*TRT01P	WEEK 8 DAY 57	Drug X	WEEK 8 DAY 57	Placebo	WORK.ADQS	-0.5127	0.7959	297	-0.64	0.5200	0.05	-2.0791	1.0537
AQLQAASC	AVISIT*TRT01P	WEEK 9 DAY 64	Drug X	WEEK 9 DAY 64	Placebo	WORK.ADQS	-1.6917	0.9454	290	-1.79	0.0746	0.05	-3.5524	0.1691

R CODES AND RELATED OUTPUTS INTERPRETATION

Firstly, fit mmrm model through the mmrm() function from package

```
library(mmrm)

library(emmeans)

library(dplyr)

fit_mmrm <- mmrm(

  formula = CHG ~ TRT01P + AVISIT + BASE + REGION1 + TRT01P*AVISIT +
us(AVISIT | USUBJID),

  data = adqs,

  reml = TRUE,

  control = mmrm_control(method = "Kenward-Roger" , vcov ="Kenward-Roger-
Linear")

)

summary(fit_mmrm)
```

In the above codes:

The formula is **CHG ~ TRT01P + AVISIT + BASE + REGION1 + TRT01P*AVISIT + us(AVISIT | USUBJID)**, it specifies response variable as CHG (change from baseline), covariates as BASE and REGION1, and **us(AVISIT | USUBJID)** specifies an unstructured covariance matrix for the time points (the visits given by AVISIT) within the subjects (the subject ID given by USUBJID).

Through **reml = TRUE**, specifies restricted maximum likelihood (REML) estimation shall be used.

Through **control = mmrm_control(method = "Kenward-Roger" , vcov ="Kenward-Roger-Linear")**, the method = Kenward-Roger and vcov = Kenward-Roger-Linear arguments specify the degrees of freedom and coefficients covariance matrix adjustment methods, respectively. Please note, if method is "Kenward-Roger" then only "Kenward-Roger" or "Kenward-Roger-Linear" are allowed for vcov.

Calling the method **summary()** will provide the covariance matrix estimate with Kenward-Roger degrees of freedom.

Part of outputs were shown as follows.

The below result is model information:

```
Formula:  CHG ~ TRT01P + AVISIT + BASE + REGION1 + TRT01P * AVISIT +
us(AVISIT | USUBJID)
Data: adqs (used 2671 observations from 400 subjects with maximum 9
timepoints)
Covariance: unstructured (45 variance parameters)
Method: Kenward-Roger
Vcov Method: Kenward-Roger-Linear
Inference: REML
```

The below table is Covariance estimate:

Table 9 Covariance estimate

	WEEK 1 DAY 8	WEEK 2 DAY 15	WEEK 3 DAY 22	WEEK 4 DAY 29	WEEK 5 DAY 36	WEEK 6 DAY 43	WEEK 7 DAY 50	WEEK 8 DAY 57	WEEK 9 DAY 64
WEEK 1 DAY 8	1.7094	-0.0941	-0.0786	0.1529	0.3339	0.0212	-0.828	-0.2445	1.6967
WEEK 2 DAY 15	-0.0941	4.2064	0.1936	-0.767	-0.2297	-0.2553	2.4416	-0.9471	-1.1111
WEEK 3 DAY 22	-0.0786	0.1936	7.6257	-1.3108	-0.6969	0.6097	0.9898	0.0287	-0.7184
WEEK 4 DAY 29	0.1529	-0.767	-1.3108	12.6084	3.0361	0.9814	1.449	-2.5107	2.5794
WEEK 5 DAY 36	0.3339	-0.2297	-0.6969	3.0361	18.1438	-0.0988	1.0155	-2.2995	2.4918
WEEK 6 DAY 43	0.0212	-0.2553	0.6097	0.9814	-0.0988	27.7966	4.2217	0.0558	-7.1588
WEEK 7 DAY 50	-0.828	2.4416	0.9898	1.449	1.0155	4.2217	41.5925	-5.0334	-0.9494
WEEK 8 DAY 57	-0.2445	-0.9471	0.0287	-2.5107	-2.2995	0.0558	-5.0334	45.6268	0.4455
WEEK 9 DAY 64	1.6967	-1.1111	-0.7184	2.5794	2.4918	-7.1588	-0.9494	0.4455	63.8735

This covariance estimate table is like SAS output **Error! Reference source not found.** Estimated R Matrix for the first subject. Although the R codes do specify the same method to estimated covariance matrix with SAS codes, here are still slight differences for some timepoint estimate, for example, in SAS output table 4 with (col9, row9) estimate is 63.8729, while in the above R table 10, (WEEK 9 DAY 64, WEEK 9 DAY 64) estimate is 63.8735. This slight difference may be caused by that unstructured covariance in SAS using the different parameterization method compare to mmrm package.

To print out fit statistics, we can split the REML criterion from the above fit model

```
Fit.Statistics <- list(
  "REML criterion" = stats::deviance(fit_mmrm),
  AIC = stats::AIC(fit_mmrm),
  AICc = stats::AIC(fit_mmrm, corrected = TRUE),
  BIC = stats::BIC(fit_mmrm)
)
```

The outputs would be like

```
 REML criterion 14759.86
 AIC 14849.86
 AICc 14851.45
 BIC 15029.48
```

The fit statistics were basically same as what derived in SAS table 6.

To obtain the least squeal means for each treatment group at each timepoint, we can library emmeans package and utilize emmeans() method, and the mmrm objects could be used with the emmeans package.

```
fit_data<- model.frame(fit_mmrm)
LSmeans_object <- emmeans::emmeans(
  fit_mmrm,
  data = fit_data,
  specs = c("AVISIT", "TRT01P"),
  weights = "proportional"
)
LSmeans_object
```

model.frame(fit_mmrm): obtains the model frame, it's actually the data frame including the variables exist in the formula.

Emmeans::emmeans(): computes estimated marginal means (EMMs) for specified factors or factor combinations in a linear model; EMMs are also known as least squares means. In the function, **fit_mmrm** is the mmrm modeling object from above mmrm codes, could be used in the emmeans function. **specs = c("AVISIT", "TRT01P")** specifies the names of the predictors over which EMMs are desired. **weights = "proportional"** specifies weight in proportion to the frequencies (in the original data) of the factor combinations that are averaged over.

The derived LSmeans_object can display the least square means (emmean), standard error, degree of freedoms for each treatment group and timepoint. Upon comparison with the SAS output table 8, it was observed that all the statistical values were essentially similar. This indicates that the mmrm object combining emmeans function is an effective tool for generating accurate statistical data.

Table 10 LSmeans estimate

AVISIT	TRT01P	emmean	SE	df	lower.CL	upper.CL
WEEK 1 DAY 8	Placebo	1.83013	0.107067	292.2989	1.61941	2.04085
WEEK 2 DAY 15	Placebo	4.062711	0.16096	309.8469	3.745998	4.379423
WEEK 3 DAY 22	Placebo	5.968103	0.217406	310.7218	5.540328	6.395878
WEEK 4 DAY 29	Placebo	7.661183	0.280283	297.1872	7.109591	8.212774
WEEK 5 DAY 36	Placebo	9.626372	0.333326	289.581	8.970324	10.28242
WEEK 6 DAY 43	Placebo	11.66987	0.40744	307.5867	10.86815	12.47159
WEEK 7 DAY 50	Placebo	13.67565	0.512648	300.1148	12.66681	14.68449
WEEK 8 DAY 57	Placebo	15.88173	0.522981	296.1731	14.8525	16.91096
WEEK 9 DAY 64	Placebo	17.99115	0.640039	289.8884	16.73143	19.25086
WEEK 1 DAY 8	Drug X	2.18338	0.111617	292.0366	1.963704	2.403056
WEEK 2 DAY 15	Drug X	3.937351	0.170679	309.6699	3.601513	4.273188
WEEK 3 DAY 22	Drug X	5.560686	0.228579	309.6728	5.110921	6.01045
WEEK 4 DAY 29	Drug X	8.217913	0.309948	297.5974	7.607945	8.827881
WEEK 5 DAY 36	Drug X	9.4389	0.385454	290.147	8.680259	10.19754
WEEK 6 DAY 43	Drug X	11.4273	0.449611	307.9297	10.5426	12.312
WEEK 7 DAY 50	Drug X	12.16312	0.550435	299.9191	11.07992	13.24632
WEEK 8 DAY 57	Drug X	15.36903	0.600015	296.7808	14.18821	16.54985
WEEK 9 DAY 64	Drug X	16.29954	0.695779	289.7783	14.93012	17.66896

To obtain the contrast between treatment group of least square means at each timepoint, we can use the emmeans package with contrast method.

Firstly, we need to create emmeans preferred visits grid list, the codes could refer as follows, it mainly based on the LSmeans_object of grid, which including the matrix of all visits and treatment groups:

```
visit_arm_grid <- LSmeans_object@grid
visit_arm_grid$index <- seq_len(nrow(visit_arm_grid))
grid_by_visit <- split(visit_arm_grid, visit_arm_grid[[vars$visit]])
fit.grid<- list()
for (j in seq_along(grid_by_visit)) {
  this_grid <- grid_by_visit[[j]]
  ref_index <- 1
```

```
  this_visit <- names(grid_by_visit)[j]
  this_ref_coefs <- numeric(nrow(visit_arm_grid))
  this_ref_coefs[this_grid$index[ref_index]] <- -1
  this_list <- list()
  for (i in seq_len(nrow(this_grid))[-ref_index]) {
    this_coefs <- this_ref_coefs
    this_coefs[this_grid$index[i]] <- 1
    this_arm <- as.character(this_grid[[vars$arm]][i])
    arm_vec <- c(arm_vec, this_arm)
    visit_vec <- c(visit_vec, this_visit)
    this_label <- paste(this_arm, this_visit, sep = ".")
    this_list[[this_label]] <- this_coefs
  }
  fit.grid <- c(fit.grid, this_list)
}
Print(fit.grid)
```

The above codes shall print the visit grid used in emmeans::contrast function, the results will be shown as follows:

```
> fit.grid
$`Drug X.WEEK 1 DAY 8`
 [1] -1  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0
$`Drug X.WEEK 2 DAY 15`
 [1]  0 -1  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0
$`Drug X.WEEK 3 DAY 22`
 [1]  0  0 -1  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0
$`Drug X.WEEK 4 DAY 29`
 [1]  0  0  0 -1  0  0  0  0  0  0  0  0  1  0  0  0  0  0
$`Drug X.WEEK 5 DAY 36`
 [1]  0  0  0  0 -1  0  0  0  0  0  0  0  0  1  0  0  0  0
$`Drug X.WEEK 6 DAY 43`
 [1]  0  0  0  0  0 -1  0  0  0  0  0  0  0  0  1  0  0  0
$`Drug X.WEEK 7 DAY 50`
 [1]  0  0  0  0  0  0 -1  0  0  0  0  0  0  0  0  1  0  0
$`Drug X.WEEK 8 DAY 57`
 [1]  0  0  0  0  0  0  0 -1  0  0  0  0  0  0  0  0  1  0
$`Drug X.WEEK 9 DAY 64`
 [1]  0  0  0  0  0  0  0  0 -1  0  0  0  0  0  0  0  0  1
```

Then, utilize emmeans::contrast method based on LSmeans_object and fit.grid list, codes refer to follows:

```
Contrast_object <- emmeans::contrast(
  LSmeans_object,
  fit.grid
)
Contrast_object
```

emmeans::contrast (): provide for contrasts, pairwise comparisons, tests, and confidence intervals based on the EMMs object. In the function, **LSmeans_object** is derived based on emmeans::eammeans() function, contains EMMs information.

The Contrast_object can display the difference of least square means (emmean), standard error, degree of freedoms for the Drug X treatment group for each timepoint. Upon comparison with the SAS output table 9, it was observed that all the statistical values were essentially similar. This indicates that the Contrast_object is an effective tool for generating accurate statistical data.

Table 11 Difference of LSmeans estimate

contrast	estimate	SE	df	t.ratio	p.value
Drug X.WEEK 1 DAY 8	0.353	0.155	293	2.279	0.0234
Drug X.WEEK 2 DAY 15	-0.125	0.235	311	-0.534	0.5939
Drug X.WEEK 3 DAY 22	-0.407	0.316	311	-1.291	0.1977
Drug X.WEEK 4 DAY 29	0.557	0.418	298	1.332	0.184
Drug X.WEEK 5 DAY 36	-0.187	0.51	290	-0.368	0.7133
Drug X.WEEK 6 DAY 43	-0.243	0.607	308	-0.4	0.6896
Drug X.WEEK 7 DAY 50	-1.513	0.752	300	-2.011	0.0453
Drug X.WEEK 8 DAY 57	-0.513	0.796	297	-0.644	0.52
Drug X.WEEK 9 DAY 64	-1.692	0.945	290	-1.789	0.0746

COMPARISON

This section shows an example of implementing MMRM both in SAS and R to generate modeling analysis outputs. In order to clarify what we found based on the above example codes and outputs interpretation, we have summarized a table for the comparison between SAS and R on MMRM analysis.

Table 12 Overall Summary on the comparison of MMRM including offset variable between SAS with R

Feature	SAS	R
Syntax	Proc mixed procedure	mmrm package to fit MMRM model, and emmeans package to compute and contrast LS means
Output format	SAS could provide outputs in a variety of formats, including SAS list, SAS datasets, HTML files, normally we can derive tables or outputs based on the SAS datasets.	The mmrm package will first generate a R S3 object, contains modeling information like sourcing data, formular, covariance estimate, etc. To drive statistics what we want, we need to combine different method, for example, summary(), model.frame() and so on. Meanwhile, mmrm model object is supported by emmeans package, so after combing emmeans::emmeans(), emmeans::contrast(), we can drive out S4 list emmGRID. Still, we don't need to look into details of emmGRID, but just print the object, we will get what we want.

Methods used in model	Both of SAS and R for MMRM, support user to specify a variety of covariance matrices, including unstructured (UN), homogeneous Toeplitz (TOEP), heterogeneous Toeplitz (TOEPH) , heterogeneous ante-dependence (ANTE(1)) , auto-regressive (AR(1)) , heterogeneous auto-regressive (ARH(1)), compound symmetry (CS) and heterogeneous compound symmetry (CSH), and fit models with restricted or standard maximum likelihood inference, perform hypothesis testing with Satterthwaite or Kenward-Roger adjustment.
Results comparison	We have provided most of the results derived from SAS Mix procedure and mmrm/emmeans package in the above sections, only very slight difference could be found in covariance estimate table, and this difference may be caused by that unstructured covariance in SAS using the different parameterization method compared to mmrm package. At the same time, for the results like model fit statistics, LS means estimation, difference of LS means estimation, the results are basically similar.

NEGATIVE BINOMIAL REGRESSION

Negative binomial regression is for modeling count variables, usually for over-dispersed count variables.

OVERVIEW OF NEGATIVE BINOMIAL REGRESSION

Regression analysis is a set of statistical processes for estimating the relationships among variables. And based on different outcome variable type, different regression model will be applied to meet the analysis or prediction need. For positive and integer outcome variables, Poisson and Negative Binomial are two most popular methods.

In statistics, Poisson regression is a generalized linear model form of regression analysis used to model count data and contingency tables. Poisson regression assumes the response variable Y has a Poisson distribution and assumes the logarithm of its expected value can be modeled by a linear combination of unknown parameters. And based on the great limitation posted on the distribution assumption, Negative Binomial Regression was generated and was considered as the popular generalization of Poisson regression because it loosens the highly restrictive assumption on equi-dispersion of Poisson distribution.

Especially when the mean of the count is lesser than the variance of the count, which is known as over-dispersion, then Negative Binomial regression is most used to test for connections between confounding and predictor variables on a count outcome variable.

ANALYSIS

After the general introduction of Negative Binomial Regression, one dummied Regional, Multicenter, Randomized, Double-Blind, Placebo Controlled, Parallel Group study, targeting to Evaluate the Efficacy of Drug X in Adults with Asthma will be analyzed by SAS and R, and the thorough comparison on coding, result and interpretation between SAS with R will be provided at last.

DATA AND ANALYSIS PURPOSE

In this section we will show an example of negative binomial regression analysis based on the dummied study. In the study, 400 subjects are randomized into Drug X and placebo, and overall asthma exacerbation count during the study period were already summarized by subject at ADaM level. The response variable in the model will be the number of asthma exacerbations experienced by a subject over the 2-year planned treatment period (AVAL), from which we explore its relationship with Treatment (TRT01PN, 1=”Drug X”, 2=”Placebo”), region (REGION1N, 1=”China”, 2=”non-China”) and history of exacerbations (EXNP12MN, 1= “<=2”, 2= “> 2”).

As assumed for a negative binomial model our response variable is a count variable, the logarithm of the time at risk (in years) for exacerbation will be used as an offset variable in the model, to adjust for

subjects having different follow-up times during which the events occur. Also, the negative binomial model, as compared to other count models (i.e., Poisson or zero-inflated models), is assumed to be the appropriate model. In other words, we assume that the dependent variable is ill-dispersed (either under- or over- dispersed) and does not have an excessive number of zeros.

SAS CODE, OUTPUTS, AND INTERPRETATION

SAS CODE

```
*Step 1 Check distribution of data;
*Subset the necessary input dataset;
data inds(keep=USUBJID TRT01P N REGION1 REGION1N EXNP12MC EXNP12MN TMATRSK
AVAL offset1);
      set adeffre(where=(paramcd="EXN" and fasfl="Y")
                   rename=(aval=_aval));
      if cmiss(TRTSDT,TRTEDT)=0 then do;
                  TMATRSK=TRTEDT-TRTSDT;*calculate time at risk;
                  offset1=log(TMATRSK/365.25); *Generate log scale of Time at
Risk for later model use;
                  end;
      if not missing(_aval) then aval=input(_aval,best.);
run;
*Check the distribution of input dataset;
proc means data=inds;
      var AVAL TMATRSK;
run;
proc freq data=inds;
      tables TRT01P*TRT01PN REGION1*REGION1N EXNP12MC*EXNP12MN/list missing;
run;
*Verify if TRT01PN is a valid expose variable;
proc sort data=inds;
      by TRT01PN;
run;
proc means data=inds;
      by TRT01PN;
      var AVAL;
run;
*Step 2 Run Negative Binomial Model;
proc genmod data=inds;
      class trt01pn region1n exnp12mn;
      model aval= trt01pn region1n exnp12mn / link=log dist=negbin
      offset=offset1 covb;
      lsmeans trt01pn / diff exp obsmargins cl;
run;
```

And here is the breakdown of above PROC GENMOD codes (SAS® Help Center, Sep2022)**:**

- **proc genmod:** is the procedure used for generalized linear regression, which including negative binomial regression also;
- **INDS:** is the input dataset, including Response variable, Predictor variables and Time at risk variable;
- **AVAL:** is the response variable;
- **TRT01PN:** is the exposure variable;
- **REGION1N, EXNP12MN:** are the covariate variables;
- **/link=log:** sets the logarithmic link function for the regression;
- **/dist=negbin:** specifies the negative binomial distribution of response variable;

- **offset=offset1:** the logarithm of the exposure time;
- **lsmeans trt01pn / diff exp obsmargins cl:** requests least squares means for the variable "**trt01pn**". The **diff** option provides differences between levels, **exp** exponentiates the estimates, **obsmargins** provides observed margins, and **cl** requests confidence limits.

SAS OUTPUTS and RESULTS INTERPRETATION

Data Distribution:

Table 13 Data Distribution of Continuous data

Variable	N	Mean	Std Dev	Minimum	Maximum
aval	400	5.5300	4.8356	1.0000	17.0000
TMATRSK	400	983.8200	205.3174	368.0000	1096.0000

Table 14 Data Distribution of Categorical data

Variable	Value	Frequency	Percent	Cumulative Frequency
TRT01PN	1	213	53.25	213
TRT01PN	2	187	46.75	400
REGION1N	1	215	53.75	215
REGION1N	2	185	46.25	400
EXNP12MN	1	211	52.75	211
EXNP12MN	2	189	47.25	400

Based on above summary statistics (**Table 13** and **Table 14**), each variable to be included in our model has 400 valid non-missing value and the mean of our response variable is lower than its variance.

Table 15 Response Variable distribution by Exposed status:

Variable	Value	N	Mean	Std Dev	Minimum	Maximum
TRT01PN	1	213	5.8028169	4.9534238	1	17
TRT01PN	2	187	5.2192513	4.691568	1	17

After by exposed status analysis (**Table 15**), the mean of response variable in each treatment group is not equal to its corresponding variance, which suggests that ill-dispersion is present and that a Negative Binomial model would be appropriate for our anlaysis. The mean for each treatment group seems different with each other while not much, thus, the relationship between response variable with exposure variable is still worth of further exporation.

Negative Binomial Regression SAS Outputs:

Table 16 The GENMOD Model Information

Data Set	INDS
Distribution	Negative Binomial
Link Function	Log
Dependent Variable	aval
Offset Variable	offset1
Number of Observations Read	400
Number of Observations Used	400

Table 17 Criteria for Assessing Goodness of Fit

Criterion	DF	Value	Value/DF
Deviance	396	414.2696	1.0461

Criterion	DF	Value	Value/DF
Scaled Deviance	396	414.2696	1.0461
Pearson Chi-Square	396	326.2577	0.8239
Scaled Pearson X2	396	326.2577	0.8239
Log Likelihood		1950.448	
Full Log Likelihood		-1106.34	
AIC (smaller is better)		2222.678	
AICC (smaller is better)		2222.831	
BIC (smaller is better)		2242.636	
Algorithm converged.			

The output begins the Model Information table and the Criteria for Assessing Goodness of Fit table. The number of observations read and used is given. There is no missing data in this dummy data, so all 400 observations that are read in are used in the analysis **(Table 16)**. In the Criteria for Assessing Goodness of Fit table, we see the Pearson Chi-Square of 326.2577 **(Table 17)**.

Table 18 Analysis of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	0.573	0.0988	0.3794	0.7666	33.65	<.0001
TRT01PN	1	1	0.1414	0.0937	-0.0423	0.3251	2.28	0.1314
TRT01PN	2	0	0	0	0	0	.	.
REGION1N	1	1	0.1091	0.0937	-0.0746	0.2928	1.35	0.2445
REGION1N	2	0	0	0	0	0	.	.
EXNP12MN	1	1	0.0531	0.0936	-0.1305	0.2366	0.32	0.5708
EXNP12MN	2	0	0	0	0	0	.	.
Dispersion		1	0.6808	0.0609	0.5713	0.8112		
Note: The negative binomial dispersion parameter was estimated by maximum likelihood.								

The Analysis of Maximum Likelihood Parameter Estimates table **(Table 18)** is presented next, which includes the regression coefficients, standard errors, the Wald 95% confidence intervals for the coefficients, chi-square tests and p-values for each of the model variables. Go into details, the variable TRT01PN has a coefficient of 0.1414, which is not statistically significant, and indicates the expected log count of the asthma exacerbation decreases by 0.1414 when **TRT01PN** change from 1 to 2. The expected log count for level 2 of **REGION1N** is 0.1091 lower than the expected log count for level 1. The expected log count for level 2 of **EXNP12MN** is 0.0531 lower than the expected log count for level 1. None of p-values of above predictor is less than 0.05, which indicates that none of above three variables is a statistically significant predictor of asthma exacerbation count.

Additionally, there is an estimate of the dispersion coefficient (often called alpha). A Poisson model is one in which this alpha value is constrained to zero. In this example, the estimated alpha has a 95% confidence interval that does not include zero, suggesting that the negative binomial model form is more appropriate than the Poisson. An estimate greater than zero suggests over-dispersion (variance greater than mean). An estimate less than zero suggests under-dispersion, which is very rare.

Table 19 Least Squares Means for TRT01PN

TRT01PN	Estimate	SE	z Value	Pr > |z|	Alpha	Lower	Upper	Exp.	Exp. Lower	Exp. Upper
1	0.801	0.06378	12.56	<.0001	0.05	0.676	0.926	2.2278	1.966	2.5245
2	0.6596	0.06865	9.61	<.0001	0.05	0.5251	0.7942	1.9341	1.6906	2.2126

Table 20 Differences of TRT01PN Least Squares Means

TRT01PN	_TRT01PN	Estimate	SE	z Value	Pr > |z|	Alpha	Lower	Upper	Exp.	Exp. Lower	Exp. Upper
1	2	0.1414	0.09372	1.51	0.1314	0.05	-0.04229	0.3251	1.1519	0.9586	1.3841

In the differences of TRT01PN least Squares Means table **(Table 20)**, estimates/SE/P value are same as the corresponding information of TRT01PN in above Analysis of Maximum Likelihood Parameter Estimates table. And the exponentiated coefficients will give the rate ratios associated with TRT01PN, indicating the multiplicative effect on the response variable. Which in this dummied data means, the rate of Asthma Exacerbation is 15% higher in Active group compared to Placebo group and the difference is not statistically significant (Exp.=1.1519, P=0.1314).

R PACKAGE, CODES, OUTPUTS, AND INTERPRETATION

R PACKAGE

The {MASS} package is a package in R that provides support functions and datasets for Venables and Ripley's book "Modern Applied Statistics with S" (4th edition, 2002). The package contains functions for regression models, classification, clustering, and more (Ripley, et al., 2023). Generally, {MASS} package will be installed along with {STAT} when R studio is installed in the first place.

R CODE

```
#Load necessary pacakges

library(MASS)

library(emmeans)

#explore distribution of target variables

summary(inds)

#Start analysis

glm_fit <- glm.nb(AVAL ~ TRT01PN + REGION1N + EXNP12MN + offset(log(TMATRSK/365.25)), data = inds , control=glm.control(maxit=1000,trace=FALSE))

#Print the model summary

summary(glm_fit)

#Check Assumption

glmpois_fit <- glm(AVAL2 ~ TRT01PN + REGION1N + EXNP12MN + offset(log(TMATRSK/365.25)), data = inds2, family = "poisson")

pchisq(2 * (logLik(glm_fit) - logLik(glmpois_fit)), df = 1, lower.tail = FALSE)

#method 1

emmeans_fit <- emmeans::emmeans(

  glm_fit,

  specs = "TRT01PN",

  data = inds2,

  type = "response",

  offset=log(1),

  weights = "proportional")

# Print the least squares means

print(emmeans_fit)

# Compute rate ratios (exponentiated coefficients)
```

```
rate_ratios <- exp(coef(glm_fit))
print(rate_ratios)
# Confidence intervals for rate ratios
conf_int <- exp(confint(glm_fit))
print(conf_int)
```

And here is the breakdown of above glm.nb and emmeans codes:

- **glm.nb():** is the function to fit a negative binomial model for data.
- **AVAL ~ TRT01PN + REGION1N + EXNP12MN + offset(log(TMATRSK/365.25)):** is the formula specified to meet our analysis purpose. **AVAL** is the response variable, **TRT01PN** is the exposure variable, **REGION1N** and **EXNP12MN** are the covariates, and the **offset()** to include the logarithm of the exposure time into our model (Year as unit).
- **data = inds:** inds is our input dataset.
- **control=glm.control(maxit=1000,trace=FALSE)):** Specify the maximum iteration times of the model as 1000.
- **emmeans()**: computes the least squares means for the target variable in the negative binomial regression model. **glm_fit** is the target negative binomial regression. **specs = "TRT01PN"** specified the names of the predictors over which EMMs are desired. **type = "response"** specifies that the inverse transformation be applied. **offset=log(1)** to ensure predictions from reference grid become rates relative to the offset that had been specified in the model, here please noted that offset() in the model has different meaning with offset statement in emmeans (Explanations supplement in emmeans package, Version 1.8.7, 2023). **weights = "proportional"** ensure the means weight in proportion to the frequencies (in the original data) of the factor combinations that are averaged over, to match the **obsmargins** in lsmeans statement in SAS code (Lenth, et al., 2023).

R OUTPUTS and RESULTS INTERPRETATION

Data Distribution:

```
> summary(inds2)
   USUBJID              TRT01P     TRT01PN       REGION1     REGION1N EXNP12MC   EXNP12MN     TMATRSK           AVAL2
 Length:400         Drug X :213   2:187   China     :215   2:185    <=2:211   2:189    Min.   : 368.0   Min.   : 1.00
 Class :character   Placebo:187   1:213   non-China:185   1:215    >2 :189   1:211    1st Qu.: 964.5   1st Qu.: 1.00
 Mode  :character                                                                     Median :1096.0   Median : 4.00
                                                                                      Mean   : 983.8   Mean   : 5.53
                                                                                      3rd Qu.:1096.0   3rd Qu.: 9.25
                                                                                      Max.   :1096.0   Max.   :17.00
```

Check model assumption:

```
> pchisq(2 * (logLik(glm_fit) - logLik(glmpois_fit)), df = 1, lower.tail = FALSE)
'log Lik.' 3.854813e-159 (df=5)
```

Summary statistics result for the input data is similar to what we got from SAS, in this dummy data, each variables to be included in our model has 400 valid non-missing value. As we mentioned earlier, negative binomial models assume the conditional means are not equal to the conditional variances which is constant in Poisson regression, thus it is suggested to use a likelihood ratio test to compare these two and test this model assumption. To do this, Poisson regression also run for our targeted formula, and above is the result. The result is far less than 0.05 and strongly suggests the negative binomial model, estimating the dispersion parameter, is more appropriate than the Poisson model.

Negative Binomial Regression R Outputs:

```
Call:
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5021  -1.3254  -0.2044   0.6688   1.4891
glm.nb(formula = AVAL2 ~ TRT01PN + REGION1N + EXNP12MN + offset(log(TMATRSK/365.25)),
    data = inds2, control = glm.control(maxit = 1000, trace = FALSE),
    init.theta = 1.468922969, link = log)
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.57299    0.09994   5.734 9.84e-09 ***
TRT01PN1     0.14139    0.09364   1.510    0.131
REGION1N1    0.10910    0.09350   1.167    0.243
EXNP12MN1    0.05309    0.09358   0.567    0.571
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1.4689) family taken to be 1)
```

```
    Null deviance: 418.24  on 399  degrees of freedom
Residual deviance: 414.27  on 396  degrees of freedom
AIC: 2222.7
2 x log-likelihood:  -2212.678
```

Model summary will cover above four Call, Deviance Residuals, Coefficients and Assessing for Goodness of Fit sections.

Detailed model information as specified in coding will occur in Call sections including the default link function also a moment estimator after an initial fit using a Poisson GLM is displayed. Additionally, R provide 5-number summary of residuals in the “Deviance Residuals” section, which showed us the summary of differences between what we observe and what our model predicts. As we can see in the report, the absolute value of Min and Max value are similar and both less than 3, and the median only slightly different from 0, which indicating the model fit fine for the data (Ford, 2022).

Then similar to SAS, R provide estimation of each coefficient, along with standard errors, z-scores, and p-values. As compared to estimation of coefficients in SAS (**Table 18**), we can see all point estimation of coefficients and the statistical significance level in the model matched, while slight difference occurred in the absolute value of standard errors and p-values, more than SAS, R added the asterisk after each p-value for highlight and easy result interpretation.

And different from SAS (**Table 17**), R provide the Assessing for Goodness of Fit section as last part of model summary reports. By comparing the result, residual deviance, AIC and log-likelihood (2 x log-likelihood: -2212.678/2=-1106.339) assessed in R got the same value as SAS outputs after digits rounding to the same.

Least Squares Means for Treatment Groups R Outputs:

```
> print(emmeans_fit)
 TRT01PN response   SE  df asymp.LCL asymp.UCL
 2          1.93 0.133 Inf      1.69      2.21
 1          2.23 0.142 Inf      1.97      2.52
```

```
Results are averaged over the levels of: REGION1N, EXNP12MN
Confidence level used: 0.95
Intervals are back-transformed from the log scale
```

Least Squares Means summary report only presents the LS means for each treatment group including corresponding standard error and confidential interval. Comparing to the SAS result (**Table 19**), based on current precision level, LS means estimation and related CI for each treatment groups are the same.

Difference in Response for Treatment Groups R Outputs:

```
> print(rate_ratios)
(Intercept)    TRT01PN1   REGION1N1   EXNP12MN1
   1.773557    1.151877    1.115272    1.054521
```

```
> print(conf_int)
                2.5 %   97.5 %
(Intercept) 1.4643997 2.157146
TRT01PN1    0.9583522 1.383953
REGION1N1   0.9278055 1.339994
EXNP12MN1   0.8774958 1.266841
```

Above **Difference in Response for Treatment Groups R Outputs** show all the exponentiated coefficients and related CI in specified model, indicating the rate ratio point estimation (exp(TRT01PN1)=1.151877) is same as what we got from SAS report (**Table 20**, Exp.=1.1519), while CI slightly different, which is expected considering the SEs of coefficients are different in the modeling result between SAS and R.

COMPARISON

Table 21 Overall Summary on the comparison of Negative Binomial analysis including offset variable between SAS with R

Level	SAS	R
Input Data	a. Offset variable need to be pre-processed in input data. b. SAS will use last level as reference level by default.	a. Offset variable can be directly added into models without pre-processing in input data. b. R will use first level as reference level by default.
Code Syntax	Used **PROC GENMOD**	Model: **glm.nb** {MASS} LS means: **emmeans** {emmeans}
Outputs	Consolidated results for model and lsmeans	Model and the follow-up analysis need to be taken care by different statistical analysis packages and functions.
Results	1. Point estimate and statistical significance level for coefficients are same, while standard errors and P values varied slightly. 2. Provide more parameters for assessing model fitness. 3. Provide CI for point estimation of coefficients in model.	1. Point estimate and statistical significance level for coefficients are same, while standard errors and P values varied slightly. 2. Only provide AIC and Deviance for assessing model fitness. 3. Provide Summary of Residual Deviance for the modeling.

To better illustrate the difference in Negative Binomial modeling between SAS with R, we would like to summarize the difference from 4 aspects as input datasets, coding syntax, outputs, and result.

First, negative binomial model requires for different input data in SAS and R, in which offset value need to be logged before putting into model in SAS, while in R log calculation could be involved in model statement.

Furthermore, as already mentioned before, procedure in SAS and package/function in R is different for the Negative Binomial regression. Considering the different procedure used in SAS and R, detailed explanation for each option within procedures should always be read and understood thoroughly so that to ensure the same analysis purpose is reached using two different software. In the meantime, some basic difference in settings of SAS and R, as sorting, default reference level, rounding, result precision level should also be kept in mind in case any unexpected mismatch.

On output level, SAS can generate and consolidate model and LS means within one PROC GENMOD procedure. While in R, glm.nb will only provide model information, if we need rate comparison between treatment groups or LS means estimation for each treatment groups then need to call emmeans in addition. On the bright side, which indicating R is more flexible than SAS, user can choose any follow-up analysis procedure as needed and preferred, while on the other hands indicating the lack of convenience in R for negative binomial regression analysis compared to SAS.

At last, from the result level, point estimation of coefficients and statistical significance levels are the same between SAS with R, while standard errors and P values varied slightly, which may occur due to variations in the default options, optimization algorithms, and underlying computations employed by the software. These differences are typically negligible and do not impact the overall interpretation or conclusions drawn from the analysis.

Overall, it's good practice to compare the key estimation outputs, such as coefficient estimates, standard errors, test statistics, and p-values, between the two software packages. For Negative Binomial regression analysis, the estimated parameter values and statistical significance should generally align. However, we should always keep in mind that if there are notable discrepancies, it is essential to review the model specifications and input data for consistency.

CONCLUSION

The comparison in MMRM and negative binomial regression based on clinical data shows that to reach the same analysis purpose as SAS, R may require combination use of multiple packages. On the one hand, indicating the flexibility of R considered that users can freely choose any package as they preferred. On the other hands, this also post some challenges for the users. First and basically, user need to be familiar with R coding syntax. Secondly, the user guide for R packages may not be thorough and detailed as SAS, which requires users to be capable to find solid reference and fix the programming errors. Lastly, understanding of the statistical method in depth is also required to better interpret the R outputs and identify the potential mistake when applying different R packages or fill in the parameters for each function.

Additionally, the output produced by R package could be the S3/S4 structured R objects, which may not be as immediately interpretable, retrievable, and easily displayed as SAS outputs. This could present a challenge for SAS programmers or statisticians who are not familiar with R languages, necessitating a greater level of proficiency in output and formatting using R language.

However, we still want to highlight the significant advantages of R language as open-source nature and powerful user-interface ability. As currently, R language is still growing at a high speed and some health authorities already started to accept the application of R in submission, we have every faith to believe other surprise may be brought by R in the near future.

As the pharmaceutical industry moves towards using different software, it is crucial to have reliable and flexible means to conduct analyses. Thus, at this transforming stage, it is a perfect timing for us to test same analysis using different statistical software. So that to better equip us with the deep understanding of statistical methods and capability of transforming to different programming language for any potential change in the future.

REFERENCES

Explanations supplement in emmeans package, Version 1.8.7. (2023). Retrieved from https://cran.r-project.org/web/packages/emmeans/vignettes/xplanations.html#offsets

Ford, C. (2022). *Understanding Deviance Residuals.* Virginia. Retrieved from https://data.library.virginia.edu/understanding-deviance-residuals/

Lenth, R. V., Bolker, B., Buerkner, P., Giné-Vázquez, I., Herve, M., Jung, M., . . . Singmann, H. (2023). *Estimated Marginal Means, aka Least-Squares Means.* Retrieved from https://github.com/rvlenth/emmeans

Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., & Firth, D. (2023). *Support Functions and Datasets for Venables and Ripley's MASS.* Retrieved from http://www.stats.ox.ac.uk/pub/MASS4/

Sabanes Bove D, D. J. (2023). *mmrm: Mixed Models for Repeated Measures. R package version 0.2.2.9027*. Retrieved from https://openpharma.github.io/mmrm/

Sabanes Bove D, D. J. (2023). *mmrm: Mixed Models for Repeated Measures. R package version 0.2.2.9027*. Retrieved from https://openpharma.github.io/mmrm/

SAS® Help Center. (Sep2022). SAS/STAT User's Guide 15.3. In *SAS® 9.4 and SAS® Viya® 3.5 Programming Documentation.* Retrieved from https://documentation.sas.com/doc/zh-CN/pgmsascdc/9.4_3.5/statug/statug_intro_sect003.htm

ACKNOWLEDGMENTS

Thanks to our manager Yiwen Wang for her great support. Also, many thanks to our colleagues Mu Zhang in providing statistical and programming support and expertise.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Xinya Wang
Xinya.wang@astrazeneca.com

Xinying Chen
Xinying.chen@astrazeneca.com