ABSTRACT

The general linear model (GLM) describes the dependent variable as a linear combination of independent variables and an error term. The GLM procedure of SAS® and the “car” package in R calculate the type I, II, or III ANOVA (analysis of variance) tables. In this study, we validated the newly-developed R package, “sasLM,” which is compatible with the GLM procedure of SAS®. The “sasLM” package was validated by comparing the output with SAS®, which is the current gold standard for statistical programming. Data from ten books and articles were used for validation. The results of the “sasLM” and “car” packages were compared with those in SAS® using 194 models. All of the results in “sasLM” were identical to those of SAS®, whereas more than 20 models in “car” showed different results from those of SAS®. As the results of the “sasLM” package were similar to those in SAS®, PROC GLM, the “sasLM” package could be a viable alternative method for calculating the type II and III sum of squares. The newly-developed “sasLM” package is free and open-source, therefore it can be used to develop other useful packages as well. We hope that the “sasLM” package will enable researchers to conveniently analyze linear models.

Keywords: Linear Model; Statistics; Analysis of Variance; Software Validation; Bioequivalence

INTRODUCTION

SAS PROC GLM stands for a “general linear model”; however, many statisticians also use the term to indicate a “generalized linear model.” While the SAS PROC GLM may be considered somewhat outdated and thus not require further support, the SAS PROC GLM is still one of the most frequently used procedures in statistics. The GLM describes the dependent variable as a linear combination of independent variables and an error term [1] and includes all three linear models: analysis of variance (ANOVA), analysis of covariance (ANCOVA), simple and multiple linear regression [2]. ANOVA is used to determine if the values of three or more unknown population means are likely to be different. Simple regression establishes the relationship between two variables using a straight line, and multiple regression explains a dependent variable using multiple continuous independent variables [3]. ANCOVA is a method that combines ANOVA and regression. One of the reasons that the GLM is frequently used is its suitability for many different types of study designs, due to
regression and ANOVA being applicable for use in experimental, quasi-experimental, and non-experimental data [4].

In ANOVA, the sum of squares (SS) is used for the hypothesis test. SAS PROC GLM provides four types of SSs;

Type I SS: Sequential SSs are calculated in the order of terms expressed in the model.
Type II SS: SSs of lower order terms are calculated adjusting terms in the same order first, then higher-order SSs are calculated adjusting terms in the same order.
Type III SS: SSs are calculated adjusting all terms (including interaction/nested terms) simultaneously.
Type IV SS: When missing cells exist, the weights of the cells are additionally adjusted.

SAS PROC GLM can be regarded as the combination of the procedures for linear model: PROC REG, PROC ANOVA, PROC TTEST, etc. It can handle both categorical and continuous variables as independent and dependent variables. The main estimation method is ‘least square method’, not ‘maximum likelihood estimation’ [5]. It only handles linear model and linear models have analytical solution for parameters and their standard errors. Therefore, it does not use ‘minimization algorithm’ nor iteration. The error is always considered as normally distributed. It does not handle link functions nor the exponential family distributions other than normal distribution. Someone who want to use ‘generalized linear model’ in SAS, he or she need to use procedures such as PROC GENMOD or PROC CATMOD. ‘Generalized linear model’ is not within the scope of this article nor “sasLM” package.

The “car” (companion to applied regression) package of R was made by Fox et al. to calculate type II or type III ANOVA tables for models [6]. Type I and II sum of squares (SSs) are more popular in the R software community. In the `anova` and `aov` functions in R, the implemented type of SSs is Type I, which is the sequential calculation. For other types of SSs, the `Anova` function from the “car” package is commonly used, which takes a `type` argument.

Traditionally, SAS GLM (or ANOVA) is used for the analysis of bioequivalence studies; however, this is not possible when using the “car” package. As such, our initial motivation for the development of the “sasLM” package was to process bioequivalence data using the following statement in R:

```
GLM(log(AUClast) ~ Sequence/Subject + Period + Treatment, data=BEdata)
```

If the above formula can be used regardless of the various crossover designs such as $2 \times 2$, $6 \times 3$, $4 \times 4$, it will be very convenient.

In this study, we validated the newly-developed R package, “sasLM,” which is compatible with the GLM procedure of SAS®. The “sasLM” package was created to implement the SSs as well as SAS® in R, a free software. Data from ten books and articles were used for validation [7-16].

**METHODS**

Although methods such as white box testing and black box testing to validate the software are available, a comparison with the gold standard (SAS®) was applied in this study.
• Validation datasets

References with datasets used for validation are as follows. The number of datasets used in the books and articles are written in parentheses.

① Harvey WR. Least-squares analysis of data with unequal subclass numbers: Agricultural Research Service, United States Department of Agriculture; 1960. (3 datasets)
② Snee RD. Computation and use of expected mean squares in Analysis of Variance. Journal of Quality Technology. 1974;6(3):128-37. (1 dataset)
③ Goodnight JH, editor. The new general linear models procedure. Proceedings of the First Annual SAS Users Group International Conference; 1976; SAS Institute Cary, NC. (4 datasets)
④ Littell RC, Stroup WW, Freund RJ. SAS for linear models: SAS Institute; 2002. (26 datasets)
⑤ Sahai H, Ojeda MM. Analysis of Variance for Random Models, Volume 2: Unbalanced Data: Theory, Methods, Applications, and Data Analysis: Springer Science & Business Media; 2004. (6 datasets)
⑥ Federer WT, King F. Variations on split plot and split block experiment designs: John Wiley & Sons; 2007. (22 datasets)
⑦ Hinkelmann K, Kempthorne O. Design and Analysis of Experiments Volume 1 Introduction to Experimental Design. 2e. John Wiley & Sons Inc. 2008.
⑧ Hinkelmann K, Kempthorne O. Design and Analysis of Experiments Volume 2 Advanced Experimental Design. John Wiley & Sons Inc. 2005. (18 datasets)
⑨ Lawson J. Design and Analysis of Experiments with SAS: CRC Press; 2010. (33 datasets)
⑩ Searle SR, Gruber MH. Linear models: John Wiley & Sons; 2016. (2 datasets)

• “sasLM” package

The “sasLM” package for the general linear model was developed in the open-source R programming language (version 3.6.3) to allow for free public use. The “sasLM” (version 0.1.4) package can be installed and loaded using the following scripts:

```r
install.packages("sasLM")
require(sasLM)
```

Detailed documentation and examples are on the online user manual in the CRAN repository (http://CRAN.R-project.org/package=sasLM), and can also be searched using “?sasLM” in the R console.

The “sasLM” package can be used by writing the following script:

```r
f8 ~ HR + SEQUENCE + PATIENT 11% SEQUENCE + VISIT + DRUG + RESIDS + RESIDT
GLM(f8, dataset)
```

• “car” package

The “car” package can be used by writing the following script (“car” version 3.0.7):

```r
install.packages("car")
require(car)
options(contrasts = c("contr.sum", "contr.poly"))
Anova(lm(f8, dataset), type = 3, singular.ok = TRUE)
```
• SAS® Software

We used the GLM procedure of SAS® version 9.4 (SAS Institute Inc., Cary, NC, USA) to obtain Type I, II, and III SSs. The SAS code is as follows:

```
CLASS PATIENT SEQUENCE VISIT DRUG RESIDS RESIDT;
MODEL HR = SEQUENCE VISIT DRUG RESIDS RESIDT PATIENT SEQUENCE+PATIENT / SSI SS2 SS3;
```

## RESULTS

The results of “sasLM,” SAS®, and “car” package were compared using 194 models. All of the results in “sasLM” showed identical results with SAS®, whereas more than 20 models in the “car” package showed different results from those of SAS® (Table 1; Supplementary Data 1). The “car” package had different outputs from the “sasLM” package or SAS® in the following models: 1 model (model 6) in the article by Snee RD [8], 1 model (model 16) in the presentation by Goodnight JH [9], 3 models (model 54, 58, 59) in the book by Littell RC [10], 2 models (model 66, 67) in the book by Sahai H [11], 10 models (model 73, 76, 78, 83, 84, 85, 87, 88, 90, 91) in the book by Federer WT [12], 1 model (model 118) in the book by Hinkelmann K [13], and 2 models (model 193, 194) in the book by Searle SR [16]. The model numbers are the same as shown in Supplementary Data 1.

One dataset that showed the same results in “sasLM,” SAS®, and “car” was in “Searle SR, Gruber MHJ, Linear Models 2e, Kindle Edition, John Wiley & Sons Inc, 2016. Example 3” [16]. The results are as follows:

• “sasLM” output

| Package | Version | Total number of models | Identical to SAS® | Different from SAS® |
|---------|---------|------------------------|-------------------|--------------------|
| sasLM   | 0.1.4   | 194                    | 194 (100%)        | 0 (0%)             |
| car     | 3.0.7   | 194                    | < 174 (90%)       | ≥ 20 (10%)         |
Validation of the R package “sasLM”

*Type III*

|                | Df | Sum Sq | Mean Sq | F value | Pr(>|F|) |
|----------------|----|--------|---------|---------|---------|
| Complexity     | 1  | 281.74 | 281.74  | 10.436  | 0.0044126 ** |
| Age            | 2  | 295.54 | 147.775 | 6.2205  | 0.0002302 ** |
| X              | 5  | 1083.68 | 216.736 | 8.0776  | 0.0001644 *** |
| Complexity:Age | 2  | 30.90  | 15.450  | 0.0707  | 0.4663711 |

---

Signif. codes: 0 ‘****’ 0.001 ‘***’ 0.01 ‘**’ 0.05 ‘*’ 0.1 ‘.’ 1

* “car” output

Anova Table (Type III tests)

|                | Sum Sq | Df | F value | Pr(>|F|) |
|----------------|--------|----|---------|---------|
| (Intercept)    | 129792 | 1  | 3376.04 | < 2e-16 *** |
| Complexity     | 352    | 1  | 10.436  | 0.0044126 ** |
| Age            | 296    | 2  | 6.1283  | 0.0088352 ** |
| X              | 1033   | 5  | 8.0776  | 0.0001644 *** |
| Complexity:Age | 39     | 9  | 0.0707  | 0.4663711 |
| Residuals      | 455    | 19 |         |          |

---

Signif. codes: 0 ‘****’ 0.001 ‘***’ 0.01 ‘**’ 0.05 ‘*’ 0.1 ‘.’ 1

* SAS® output

**Searle Example 3**

The GLM Procedure

Dependent Variable: Y

| Source               | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|----------------------|----|----------------|-------------|---------|--------|
| Model                | 10 | 1465.555555    | 1465.555555 | 6.07    | 0.0004 |
| Error                | 19 | 458.607110     | 24.197216   |         |        |
| Corrected Total      | 29 | 1924.162667    |             |         |        |

R-Square | Coef Var | Root MSE | Y Mean
---|----------|----------|---------|
0.761669 | 0.073829 | 4.912964 | 98.83333

| Source               | DF | Type I SS | Mean Square | F Value | Pr > F |
|----------------------|----|-----------|-------------|---------|--------|
| Complexity           | 1  | 67.500000 | 67.500000   | 2.93    | 0.1199 |
| Age                  | 2  | 313.446526 | 156.723164 | 6.49    | 0.0071 |
| X                    | 5  | 1045.624566 | 209.12911 | 8.56    | 0.0002 |
| Complexity:Age       | 2  | 38.988673  | 19.494336  | 0.81    | 0.4606 |

| Source               | DF | Type II SS | Mean Square | F Value | Pr > F |
|----------------------|----|------------|-------------|---------|--------|
| Complexity           | 1  | 241.717590 | 241.717590 | 10.01   | 0.0051 |
| Age                  | 2  | 300.904217 | 150.45208  | 6.23    | 0.0083 |
| X                    | 5  | 1003.476223 | 200.695245 | 8.99    | 0.0002 |
| Complexity:Age       | 2  | 38.988673  | 19.494336  | 0.81    | 0.4606 |

| Source               | DF | Type III SS | Mean Square | F Value | Pr > F |
|----------------------|----|-------------|-------------|---------|--------|
| Complexity           | 1  | 251.733138 | 251.733138 | 10.48   | 0.0044 |
| Age                  | 2  | 255.533075 | 127.766385 | 6.13    | 0.0038 |
| X                    | 5  | 1083.478223 | 216.696445 | 8.88    | 0.0002 |
| Complexity:Age       | 2  | 38.988673  | 19.494336  | 0.81    | 0.4606 |
One dataset that showed the same results in “sasLM” and SAS® but not in “car” was in “Searle SR, Gruber MH, Linear Models 2e, Kindle Edition, John Wiley & Sons Inc., 2016, Page 390” [16]. The “car” result was different from others because the model had aliased coefficients. The results are as follows:

- “sasLM” output

```r
$ANOVA
Response : weight

                 DF Sum Sq Mean Sq F value Pr(>F)
MODEL           7   82.11 11.714  2.0916 0.14
RESIDUALS      10   56.00  5.600
CORRECTED TOTAL 17  138.00

$`Type I`

                 DF Sum Sq Mean Sq  F value Pr(>F)
treatment       2   10.50  5.250 0.9375 0.42345
variety         3  36.786 12.262 2.1896 0.18233
interaction     2  34.714 17.357 3.0955 0.08965

Signif. codes:  *** 0.001 ** 0.01 * 0.05 . 0.1 1

$`Type II`

                 DF Sum Sq Mean Sq  F value Pr(>F)
treatment       2   9.486  4.743 0.8469 0.45731
variety         3  36.706 12.235 2.1856 0.15232
interaction     2  34.714 17.357 3.0995 0.08965

Signif. codes:  *** 0.001 ** 0.01 * 0.05 . 0.1 1

$`Type III`

                 DF Sum Sq Mean Sq  F value Pr(>F)
treatment       2  12.471  6.235 1.1134 0.36555
variety         3  34.872 11.624 2.0757 0.16719
interaction     2  34.714 17.357 3.0955 0.08965

Signif. codes:  *** 0.001 ** 0.01 * 0.05 . 0.1 1

- “car” output

Note: model has aliased coefficients
sums of squares computed by model comparison
Anova Table (Type III tests)

Response: weight

                  Sum Sq DF F value Pr(>F)
treatment      0.000  0
variety        0.000  0
interaction   34.714  2 3.0995 0.08965
Residuals      56.000 10

Signif. codes:  *** 0.001 ** 0.01 * 0.05 . 0.1 1
DISCUSSION

We developed the “sasLM” package for R for the analysis of linear models. Type I, II, and III SSs can be obtained by using the “sasLM” the same way as in SAS®. As the results of this package are identical to the SAS® PROC GLM (or ANOVA or REG), the “sasLM” could be a viable alternative method for calculating the SSs.

There can be slight differences (off by one in the last digit) among the output. This results from the round-to-even number way of the R rounding function, which affects all R packages including “sasLM.” Unlike Type II SSs, Type III SSs simultaneously adjusts the interaction effect with the first-order primary effects. While type I or II SSs are preferred over type III SSs in some cases [17], the type III SSs is under considerable demand but was not available in R.

There are several reasons for the difference between “sasLM” and “car” in some validation datasets. SSs for nesting or whole plot factors in the nested design or split-plot design are not calculated by “car” package. When the degree of freedom of residual is 0, “car” does not produce the output. When there are aliased coefficients, the “car” package calculates SSs by the model comparison method, while SAS calculates SSs using g2 generalized inverse [9,10].
Although the “sasLM” can also be used to analyze the bioequivalence data, our current opinion is that tools originally intended for linear mixed-effects models (e.g., PROC MIXED, nlme) are superior to those originally intended for fixed-effects models (e.g., GLM, ANOVA, sasLM). Despite this drawback, there are still infinite designs of experiments using fixed effects. Therefore, we believe that the “sasLM” package can be useful in many cases.

The “sasLM” package was validated by comparing the model outputs with SAS® PROC GLM because the current gold standard for the statistical program is SAS®. Software validation is carried out to determine whether the right product is being built. This is a dynamic process of testing to ensure that the software meets its intended use when supplied. Although there are many ways to validate a software [17,18], we thought that it is best to compare with the current gold standard [19].

The newly-developed “sasLM” package is free and open-source, so it can be used to develop other useful packages as well. We hope that the “sasLM” package will enable researchers to conveniently analyze linear models.

SUPPLEMENTARY MATERIAL

Supplementary Data 1
Validation of ‘sasLM’ Package

Click here to view

REFERENCES

1. Kiebel S, Holmes A. The general linear model. In: Frackowiak RSJ, Friston KJ, Firth CD, Dolan RJ, Mazziotta JC (eds). Human brain function. 2nd ed. London: Academic Press; 2003, 725-60.
2. Miller J, Haden P. Statistical analysis with the general linear model. 2006.
3. Myers RH, Myers RH. Classical and modern regression with applications. Belmont (CA): Duxbury Press; 1990.
4. Rutherford A. Introducing ANOVA and ANCOVA: a GLM approach. New York (NY): Sage; 2001.
5. SAS Institute. Base SAS 9.4 procedures guide. Cary (NC): SAS Institute; 2015.
6. Fox J, Weisberg S, Adler D, Bates D, Baud-Bovy G, Ellison S, et al. Package ‘car’. Vienna: R Foundation for Statistical Computing; 2012.
7. Harvey WR. Least-squares analysis of data with unequal subclass numbers. Washington, D.C.: Agricultural Research Service, United States Department of Agriculture; 1960.
8. Snee RD. Computation and use of expected mean squares in analysis of variance. J Qual Technol 1974;6:128-137.
9. Goodnight JH. The new general linear models procedure. In: Proceedings of the First Annual SAS Users Group International Conference. Cary (NC): SAS Institute; 1976.
10. Littell RC, Stroup WW, Freund RJ. SAS for linear models 4th ed. Cary (NC): SAS Institute; 2002.
11. Sahai H, Ojeda MM. Analysis of variance for random models, volume 2: unbalanced data: theory, methods, applications, and data analysis. Berlin: Springer Science & Business Media; 2004.
12. Federer WT, King F. Variations on split plot and split block experiment designs. Hoboken (NJ): John Wiley & Sons; 2007.
13. Hinkelmann K, Kempthorne O. Design and Analysis of Experiments Volume 1 Introduction to Experimental Design. 2e. John Wiley & Sons Inc. 2008.

https://tcpharm.org
https://doi.org/10.12793/tcp.2020.28.e9

90
14. Hinkelmann K. Design and analysis of experiments, volume 3: special designs and applications. Hoboken (NJ): John Wiley & Sons; 2012.
15. Lawson J. Design and analysis of experiments with SAS. Boca Raton (FL): CRC Press; 2010.
16. Searle SR, Gruber MH. Linear models. Hoboken (NJ): John Wiley & Sons; 2016.
17. Sommerville I. Software engineering, 9th ed. London: Pearson; 2011.
18. Howden WE. Applicability of software validation techniques to scientific programs. ACM Trans Program Lang Syst 1980;2:307-320.
19. Jørgensen M. Relative estimation of software development effort: it matters with what and how you compare. IEEE Softw 2012;30:74-79.