SESUG Paper 36-2025

Navigating Model Choice in SAS: A Case-Based Decision-Support Framework for PROC REG, GLM, GENMOD, and MIXED

Sriya Venkat, North Carolina State University
Jonathan Duggins, North Carolina State University

Abstract

SAS provides a wide range of procedures for modeling continuous, binary, and count data. In this paper, we will analyze the uses of PROC REG, PROC GLM, PROC GENMOD, and PROC MIXED. These procedures, though they sometimes have overlaps in output, significantly differ in their methods of estimation, complexity, and underlying assumptions. This paper builds on previous comparative work and develops a practical guide for biostatisticians working with real-world health data. Using examples from various clinical trials and epidemiological studies, we demonstrate how estimation, interpretability, and fit are impacted by model choice - especially when we have causes of non-independence, non-normality, and repeated measures. We evaluate these four procedures on case studies involving blood pressure monitoring, adverse event counts, and longitudinal biomarker data. The result is a decision-support framework that helps researchers select and apply the most appropriate procedure based on study design, outcome type, and expected modeling outcomes. This crosswalk aims to support training and transparency for analysts in the health sciences using SAS for complex data structures.

Introduction

In health sciences research, modeling choices are extremely important. The same dataset can yield different results depending on the statistical procedure that is used. Those differences often stem from the assumptions that are built into each model. When analysts and researchers treat categorical predictors as numeric, they ignore correlation between repeated measures, or apply the wrong distribution to binary outcomes, and they risk drawing incorrect conclusions. In this paper, we directly compare four SAS procedures:

- `PROC REG`: Used for standard linear regression;
- `PROC GLM`: Used for general linear models with CLASS effects;
- `PROC GENMOD`: Used for generalized linear models (e.g., logistic, binary);
- `PROC MIXED`: Used for linear mixed-effects models for correlated data.

We illustrate the application of each procedure using publicly available data from the **National Health and Nutrition Examination Survey (NHANES)**, 2017–2018 cycle. This paper is broken down into two case studies:

1. **Case Study 1 – Continuous Outcome (Systolic Blood Pressure):** Comparison of `PROC REG`, `PROC GLM`, and `PROC MIXED` for modeling continuous outcomes.
2. **Case Study 2 – Binary Outcome (High Blood Pressure Status):** Application of `PROC GENMOD` for modeling binary outcomes.

For each case study, we outline the research question, data preparation steps, statistical models, and mathematical foundations. We also interpret results in context and explain why each procedure produces the results it does.

Case Study 1 - Continuous Outcome with Systolic Blood Pressure

Blood pressure is a clinically important outcome and treatment decisions often depend on whether mean blood pressure differs between groups. In this case study, we use simulated treatment groups within NHANES data to illustrate the process of testing whether mean systolic blood pressure varies significantly across three treatments. Specifically answering this question provides a clear framework to compare and contrast modeling approaches.

In this first case study, we investigated continuous systolic blood pressure (SBP) data from the National Health and Nutrition Examination Survey (NHANES) 2017–2018 cycle. During a single examination session, each participant in NHANES could have up to four valid systolic readings, which are labeled `BPXSY1`–`BPXSY4`. Along with the SBP information for each participant, we obtained demographic information, including age and sex, from the corresponding demographics file.

Because the dataset we obtained from NHANES was purely observational, we created a simulated variable, `Treatment`, assigning participants randomly to one of three groups - Placebo, DrugA, or DrugB - to mimic the structure of a controlled experiment. To ensure reproducibility, we fixed a random seed and assigned treatments using a uniform random number generator in SAS:

Program 1: Creating Simulated Treatment

```
data casestudy1;
    set casestudy1;
    call streaminit(12345);
    r = rand("Uniform");
    if r < 1/3 then Treatment = "Placebo";
    else if r < 2/3 then Treatment = "DrugA";
    else Treatment = "DrugB";
    drop r;
run;
```

After creating our similated Treatment variable, we then merged the demographic file and blood pressure file by participant identifier (`SEQN`). We calculated the mean systolic blood pressure for each participant across all valid readings using the `MEAN` function in SAS to simplify the analysis and avoid handling multiple individual values:

Program 2: Creating Simulated Treatment

```
data nhanes_sbp_avg;
    set nhanes_bp_data;
    Mean_SBP = mean(of BPXSY1 BPXSY2 BPXSY3 BPXSY4);
run;
```

Through these steps, we obtained a baseline dataset containing participant ID, sex, age, the simulated treatment group, and mean systolic blood pressure, which we refer to as `casestudy1`.

To analyze the benefits and consequences of model choice, we fit three models to these data. First, we attempted to use `PROC REG`, with the treatment variable entered directly into the model:

Program 3: Creating Simulated Treatment

```
proc reg data=casestudy1;
    model Mean_SBP = Treatment;
    title "PROC REG: Treating Treatment as Continuous (Incorrect Model)";
run;
quit;
```

When this code is executed, SAS produces the following error:

SAS returns the following error and warnings:

```
ERROR: Variable Treatment in list does not match type prescribed for this list.
NOTE: The previous statement has been deleted.
WARNING: No variables specified for an SSCP matrix. Execution terminating.
NOTE: PROCEDURE REG used (Total process time):
      real time           0.19 seconds
      cpu time            0.04 seconds
```

This error arises because `Treatment` is a categorical variable with three distinct levels, all of which are unordered: *Placebo*, *DrugA*, and *DrugB*. `PROC REG`, by default, requires all predictors to be numeric and interprets them as continuous variables. Since `Treatment` is a character variable, SAS will not execute the code as written unless the values are manually recoded to numeric form. Recoding the values would look like this:

- Placebo $\rightarrow$ 0
- DrugA $\rightarrow$ 1
- DrugB $\rightarrow$ 2

However, doing this introduces an unjustified assumption of ordinality - namely, that there is some type of linear progression in the outcome from *Placebo* to *DrugA* to *DrugB*. In this case, the fitted model in `PROC REG` would be:

$$\hat{Y}_i = \beta_0 + \beta_1 T_i + \varepsilon_i$$

where:

- $\hat{Y}_i$ is the predicted mean SBP for participant i
- T_i is the numeric code for the treatment group (0, 1, 2)
- β_0 is the intercept (predicted mean SBP for *Placebo*, where $T_i = 0$)
- β_1 is the estimated change in mean SBP per one-unit increase in treatment code

Having this parametrization forces the following upon us:

$$\mu_{\text{DrugA}} - \mu_{\text{Placebo}} = \mu_{\text{DrugB}} - \mu_{\text{DrugA}} = \beta_1$$

which means that the difference between *Placebo* and *DrugA* must exactly be equal to the difference between *DrugA* and *DrugB*. This imposed restriction is very rarely valid for treatment categories that are independent. If the true group differences are non-monotonic or unequal (e.g., *DrugA* $<$ *Placebo* but *DrugB* $>$ *Placebo*), the model is mis-specified and β_1 will be biased.

Could PROC REG be made to work?

The previous analysis of the modeling flaw in PROC REG brings us to the question - Could we make PROC REG work? In short, yes, we could, but only if the categorical predictor is transformed into dummy variables (0/1 indicators) representing each treatment group relative to a baseline. For example, using *Placebo* as the reference category:

$$D_1 = \begin{cases} 1 & \text{if DrugA} \\ 0 & \text{otherwise} \end{cases} \qquad D_2 = \begin{cases} 1 & \text{if DrugB} \\ 0 & \text{otherwise} \end{cases}$$

The model then becomes:

$$\hat{Y}_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \varepsilon_i$$

Here, we see that β_1 is the estimated difference in mean SBP between *DrugA* and *Placebo*, and β_2 is the estimated difference between *DrugB* and *Placebo*. Mathematically, this is equivalent to the model produced by `PROC GLM` when `Treatment` is declared as a `CLASS` variable. However, when we use `PROC REG` in this way, manual creation of dummy variables must be done, which greatly increases time spent, and the risk of coding errors and inconsistencies.

For categorical predictors, especially those with more than two levels, procedures such as `PROC GLM` are preferred, as they handle dummy coding internally, ensure correct parameterization, and provide least squares means and pairwise comparisons without additional data manipulation.

The limitations of `PROC REG` in handling categorical predictors highlight the importance and efficiency of using a procedure specifically designed for such variables. To correctly model treatment effects when the predictor is categorical, we turn to the General Linear Model via `PROC GLM`.

PROC GLM: Correctly Modeling Categorical Treatment Groups

We applied `PROC GLM`, specifying `Treatment` as a `CLASS` variable so SAS would treat it as categorical. The corresponding SAS code was:

Program 4: Creating Simulated Treatment

```
proc glm data=casestudy1;
    class Treatment;
    model Mean_SBP = Treatment;
    lsmeans Treatment / stderr pdiff cl;
    title "PROC GLM: Categorical Treatment Comparison (Correct Model)";
run;
quit;
```

Contrast to `PROC REG`, `PROC GLM` automatically performs dummy (indicator) coding for categorical predictors. When `Treatment` is listed in the `CLASS` statement, SAS internally creates $k-1$ dummy variables for k categories, using one of the groups as the reference group (by default, the last in alphabetical order unless otherwise specified, or in this case, Placebo).

In most experimental or clinical trial analyses, the Placebo group is the baseline group against which all other treatments are compared.

When we set *Placebo* as the reference category, the model becomes:

$$\hat{Y}_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \varepsilon_i$$

where:

- $\hat{Y}_i$ is the predicted mean SBP for participant i
- $D_{1i} = 1$ if participant i is in *DrugA*, 0 otherwise
- $D_{2i} = 1$ if participant i is in *DrugB*, 0 otherwise
- β_0 = mean SBP for the *Placebo* group
- β_1 = difference in mean SBP between *DrugA* and *Placebo*
- β_2 = difference in mean SBP between *DrugB* and *Placebo*
- $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ = residual error

Why PROC GLM is Appropriate Here

Unlike `PROC REG`, `PROC GLM` does not impose a false linear ordering among treatment groups. Instead, it allows each treatment mean to be estimated independently relative to the baseline, so:

$$\mu_{\text{Placebo}},\ \mu_{\text{DrugA}},\ \mu_{\text{DrugB}}$$

can each take on any value, with no constraints that differences must be equal or monotonic. As in PROC REG, with GLM, parameter estimates are obtained via ordinary least squares (OLS):

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

where $\mathbf{X}$ is the design matrix containing the intercept and dummy variables for the non-reference treatment groups.

LSMEANS and Pairwise Comparisons

The `LSMEANS` statement in SAS computes *least squares means*. LSMEANS, unlike raw group means, are derived from the fitted model and account for other variables or imbalances in the data. This makes them especially useful in models where the design may be unbalanced or where covariates are present, because they provide a fair, model-adjusted estimate of each group's mean outcome. For our GLM, we used LSMEANS to estimate adjusted group means for each treatment, control for the model structire, and enable direct comparisons between treatments through pairwise tests. In the context of this study, LSMEANS provides unbiased, model-adjusted treatment means rather than simple raw averages, which could be misleading in unbalanced designs, and PDIFF enables us to directly answer the question: Are any treatments significantly different from one another?

For our model, the LSMEANS are expressed in terms of the estimated regression coefficients:

$$\text{LSMean}_{\text{Placebo}} = \hat{\beta}_0$$

$$\text{LSMean}_{\text{DrugA}} = \hat{\beta}_0 + \hat{\beta}_1$$

$$\text{LSMean}_{\text{DrugB}} = \hat{\beta}_0 + \hat{\beta}_2$$

where $\hat{\beta}_0$ is the intercept (mean for Placebo), and $\hat{\beta}_1$ and $\hat{\beta}_2$ are the estimated effects of DrugA and DrugB compared to Placebo.

Pairwise Comparisons (PDIFF option)

We included the `PDIFF` option to test whether the differences between LSMEANS for any two treatment groups are statistically significant. The null hypothesis for each test is:

$$H_0 : \mu_g = \mu_{g'}$$

For example: Placebo vs. DrugA, Placebo vs. DrugB, and DrugA vs. DrugB.

These tests are t-tests of the form:

$$t = \frac{\hat{\mu}_g - \hat{\mu}_{g'}}{SE(\hat{\mu}_g - \hat{\mu}_{g'})}$$

where the standard error comes from the model's residual variance, and degrees of freedom are based on the GLM's error term.

By combining LSMEANS and PDIFF, we ensured our conclusions about treatment differences were statistically rigorous and based on adjusted, not raw, comparisons.

Interpretation for Our Data

The LSMEANS for DrugA, DrugB, and Placebo were almost identical. All pairwise p-values exceeded 0.05, which indicates no statistically significant differences between treatments. The plot of LSMEANS showed overlapping 95% confidence intervals, visually supporting the statistical results. The hypothesis tests confirm the lack of statistically significant differences, rather than the overlap of confidence intervals alone.

This suggests that:

$$\mu_{\text{Placebo}} \approx \mu_{\text{DrugA}} \approx \mu_{\text{DrugB}}$$

Given that treatment assignments were randomly simulated with no true intervention effect, these findings are consistent with expectations.

When to Use PROC GLM Instead of PROC REG

From conducting this case study thus far, we can, with certainty, say that `PROC GLM` (or `PROC MIXED` for more complex designs) should be used instead of `PROC REG` when:

- The predictor is categorical with $k > 2$ levels
- Pairwise comparisons or adjusted means are needed
- Automatic dummy coding is preferred to avoid manual recoding errors
- The design is balanced or unbalanced but still meets general linear model assumptions

Moreover, `PROC GLM` can accommodate models with multiple categorical and continuous predictors, interactions, and factorial designs - making it far more flexible for ANOVA-type analyses.

PROC MIXED: Correctly Handling Repeated SBP Readings Within Participants

In the PROC GLM analysis, each participant's mean systolic blood pressure (SBP) was treated as a single, independent observation. This ignores the fact that the NHANES dataset contains up to four SBP readings per participant (BPXSY1–BPXSY4), and repeated readings from the same person are *not* independent—people with high SBP on one reading tend to have high SBP on the others.

If this within-person correlation is ignored, as in a basic regression or PROC GLM, the model can underestimate standard errors and overstate statistical significance (inflating Type I error rates). **PROC MIXED** addresses this problem by:

- allowing multiple observations per participant,
- accounting for within-person correlation with a participant-specific random intercept, and
- using all available readings, even if some are missing.

Data Restructuring (Wide $\rightarrow$ Long)

We reshaped each participant's four SBP columns into a long format so each row corresponds to a single reading (SEQN, Treatment, Visit, SBP), using `PROC TRANSPOSE`.

Model Specification

We fit a *random-intercept linear mixed model* to estimate treatment effects while accounting for repeated measurements from the same participant. For participant i and reading j:

$$SBP_{ij} = \beta_0 + \tau_1 I(\text{DrugA}_{ij}) + \tau_2 I(\text{DrugB}_{ij}) + b_i + \varepsilon_{ij},$$

where β_0 is the placebo mean, τ_1 and τ_2 are treatment effects (DrugA/DrugB vs. placebo), $I(\cdot)$ indicates group membership, $b_i \sim N(0, \sigma_b^2)$ is a participant-specific random intercept, and $\varepsilon_{ij} \sim N(0, \sigma^2)$ is measurement error.

SAS Implementation

```
proc mixed data=bp_long method=reml;
    class SEQN Treatment Visit;
    model SBP = Treatment / solution ddfm=satterth;
    random intercept / subject=SEQN;
    lsmeans Treatment / pdiff cl;
    title "PROC MIXED: Random Intercept Model for Repeated Systolic BP";
run;
```

Why MIXED is Preferred over GLM or Standard Regression

By default, **GLM/regression** assumes independent observations and therefore ignores correlation among repeated readings from the same person, yielding anti-conservative inference. **MIXED** explicitly models the clustering (participant-level random intercept), producing valid standard errors and tests for treatment effects.

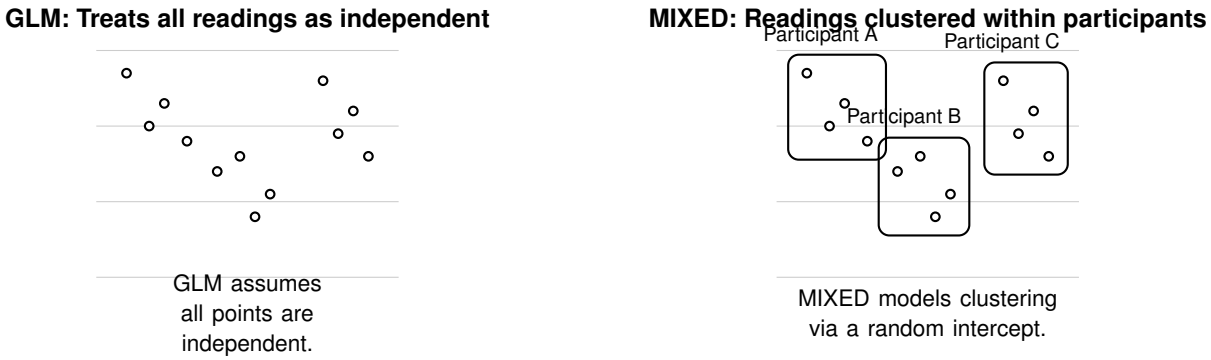

Figure 0.1: Conceptual contrast between GLM and MIXED for repeated measures.

Results

The `PROC REG` analysis demonstrated the limitation of trying to model a categorical variable as if it were numeric. The output gives a single slope for `Treatment`, which would imply a straight-line increase or decrease in SBP as you move from one coded treatment value to the next. This does not make sense for three unrelated treatment groups.

In contrast, the `PROC GLM` analysis treated `Treatment` correctly as a categorical variable, producing separate least squares means for each group. The GLM output showed that the mean SBP values for *DrugA*, *DrugB*, and *Placebo* were all very close to each other. The pairwise comparison p-values were all greater than 0.05, meaning there were no statistically significant differences between any of the groups. This is expected, because the treatment groups in this dataset were assigned randomly and do not reflect real interventions.

When repeated measures were incorporated using `PROC MIXED`, the conclusions stayed the same: the mean SBPs were still very similar across groups, and none of the differences were statistically significant. The advantage of the mixed model is that it used all individual readings instead of per-person averages and adjusted the standard errors to account for the correlation among repeated measurements for the same participant. This gave a more appropriate error structure and demonstrated how to handle repeated-measures data correctly, even though in this case the treatment effect itself was not significant.

Case Study 2: Modeling a Binary Health Outcome with GENMOD

In Case Study 1, we modeled a continuous outcome (systolic blood pressure) using `PROC MIXED` to handle correlated repeated measures within individuals. However, not all outcomes in health data are continuous - many are binary, such as the presence or absence of a disease. In these cases, the assumptions of linear models no longer hold, and specialized methods are needed. To illustrate this, we turn to `PROC GENMOD`, which fits generalized linear models appropriate for binary or count outcomes.

Background

High blood pressure (hypertension) is a clinically important binary outcome - patients are either classified as having high blood pressure or not. Understanding whether the probability of high blood pressure differs

between population subgroups can provide valuable epidemiologic insights. In this case study, we investigate whether the probability of high blood pressure differs between males and females using self-reported data from the National Health and Nutrition Examination Survey (NHANES) 2017–2018 cycle.

Data Preparation

High blood pressure status was obtained from the blood pressure questionnaire file `BPQ_J.xpt`, specifically variable `BPQ020` (*"Have you ever been told by a doctor or other health professional that you had high blood pressure?"*). Responses of "Yes" were coded as 1 and "No" as 0, creating the binary outcome variable `High_BP`. All other responses, including missing values, "Don't know," and refusals, were excluded.

Demographic information, including gender (`RIAGENDR`), was obtained from the demographics file `DEMO_J.xpt`. The two datasets were merged by participant identifier (`SEQN`). The resulting dataset contained `SEQN`, gender, and high blood pressure status for analysis.

Model Specification

To evaluate the effect of gender on the probability of high blood pressure, we used `PROC GENMOD` to fit a logistic regression model with a logit link function. Gender was modeled as a categorical predictor. The model can be expressed as:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \cdot I(\text{Male}_i)$$

where:

- π_i = probability that participant i has high blood pressure,
- $I(\text{Male}_i)$ = 1 if male, 0 if female (reference group),
- β_0 = log-odds of high blood pressure for females,
- β_1 = log-odds ratio comparing males to females.

The odds ratio for males versus females is given by $\exp(\beta_1)$.

SAS implementation:

```
proc genmod data=nhanes_q_ready;
    class RIAGENDR;
    model High_BP = RIAGENDR / dist=binomial link=logit;
    lsmeans RIAGENDR / ilink cl oddsratio;
    title "PROC GENMOD: High Blood Pressure by Gender";
run;
```

The `LSMEANS` statement with the `ILINK` option back-transforms estimates from the log-odds scale to predicted probabilities, while the `ODDSRATIO` option produces interpretable effect size measures.

Results

The GENMOD analysis estimated the probability of high blood pressure to be **0.6663** (95% CI: 0.6421, 0.6900) for females and **0.5928** (95% CI: 0.5654, 0.6197) for males. The odds ratio for males versus females was **0.76** (p-value > 0.05), indicating no statistically significant gender difference. The wide overlap of the confidence intervals along with the non-significant p-value provide consistent evidence that any apparent difference between males and females is likely due to sampling variability rather than a true effect.

When to Use PROC GENMOD

`PROC GENMOD` is best suited for generalized linear models where:

- The outcome follows a distribution in the exponential family (e.g., binomial for binary outcomes, Poisson for counts).
- You want to use link functions (logit, log, identity) to relate predictors to the mean of the outcome.
- Data are independent observations, or correlation is modeled using generalized estimating equations (GEE).
- You do not require subject-specific random effects.

Compared to the other procedures in this paper:

- `PROC GLM` handles only continuous outcomes with normally distributed errors.
- `PROC MIXED` extends to correlated or clustered data for continuous outcomes.
- `PROC GLIMMIX` extends mixed modeling to non-normal outcomes, including binary, with random effects.
- `PROC GENMOD` is ideal when the outcome is binary or count data, there are no random effects, and you want *population-averaged* interpretations rather than subject-specific effects.

Conclusion

This paper compared four SAS procedures -`PROC REG`, `PROC GLM`, `PROC MIXED`, and `PROC GENMOD` - through two case studies using NHANES data. With each case, we illustrated the importance of aligning model choice with study design and outcome type. Each case demonstrated the importance of aligning model choice with study design and outcome type.

In Case Study 1, we examined systolic blood pressure as a continuous outcome with simulated treatment groups. `PROC REG`, when used naively, demonstrated why categorical predictors cannot be treated as continuous without changing the interpretation. `PROC GLM` addressed this limitation by handling dummy coding internally and producing valid least squares means and pairwise comparisons. `PROC MIXED` further extended the analysis by appropriately incorporating repeated measures, using all individual blood pressure readings rather than collapsing them into per-person averages, and accounting for within-subject correlation through random effects.

In Case Study 2, we turned to a binary outcome - self-reported high blood pressure - and modeled it as a function of gender. `PROC GENMOD` provided estimates of probabilities, confidence intervals, odds ratios, and hypothesis tests under a logistic regression framework. The analysis revealed no statistically significant differences between males and females, which was expected given the observational nature of the data.

Across both studies, our findings reinforce three major lessons. First, improper use of procedures (e.g., `PROC REG` with categorical predictors) can lead to misleading results. Second, procedures designed for the correct data structure (`GLM` for continuous outcomes, `GENMOD` for binary outcomes, `MIXED` for repeated measures) provide interpretable, unbiased estimates and valid inference. Third, model choice impacts not only statistical results but also interpretability and reproducibility, both of which are pillars of applied biostatistics.

Ultimately, these case studies illustrate how estimation, interpretability, and fit are each influenced by the choice of modeling procedure. By contrasting approaches across continuous, binary, and repeated-measures outcomes, we provide a framework for selecting the most appropriate SAS procedure in practice. This crosswalk can aid both training and applied research, helping analysts in the health sciences to implement models that match the structure of their data and the goals of their study.

Recommended Readings

References

[1] Eloise Barry/London. Scotland Just Showed How Easy It Is to End 'Period Poverty.' *Time*, 15 Aug. 2022. https://time.com/6206216/scotland-law-period-poverty/?utm_source

[2] Yarandi, H. N. Comparison of PROC MIXED and PROC GLM for Analysis of Repeated Measures Data. In *Proceedings of the SAS Users Group International Conference (SUGI 31)*, 2006. Available at: file://wolftech.ad.ncsu.edu/cos/stat/redirect/jwduggin/Downloads/ComparisonofPROCMIXEDandPROCGLMforAnalysisofRepeatedMeasuresDataSD06-Yarandi.pdf

[3] Little, R., Stroup, W., & Freund, R. (2002). *SAS for Linear Models*. Cary, NC: SAS Institute Inc.

Contact Information

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

Sriya Venkat
North Carolina State University
srvenka2@ncsu.edu

Jonathan Duggins
North Carolina State University
jwduggin@ncsu.edu
https://jonathanduggins.com/