# Introduction to SAS procedure

### PROC TABULATE

: Create descriptive statistics in table format, such as mean, standard deviation, median, quartile, percentage, minimum, maximum, and so on.

| Function | Statement |
|---|---|
| Var | Numeric format only. Used to generate summary statistics. |
| Class | Classify the state of the variables. |
| Table | Specify the expression for the row and column. |
| By | Make a table for each subgroup. |
| Freq | Identify a variable that indicates the frequency of each value. |

### PROC SGPLOT

: Create visualizations to represent statistics using scatter plots, line plots, histograms, and regression plots

### PROC GLM

: To fit general linear models, the GLM procedure uses the least squares method. The PROC GLM procedure may be used for a variety of studies, including simple/multiple regression, analysis of variance (ANOVA), analysis of covariance, multivariate analysis of variance (MANOVA), and repeated measures analysis of variance.

| Function | Statement |
|---|---|
| Class | Classify the variables status |
| Model | The model that will be fit general linear models. |
| By | Variables to make subgroups for the analysis. |
| Estimate | Estimates the parameters′ linear functions.. |
| Means | Calculates and compares arithmetic means. |
| Lsmeans | Calculates least squares means. |
| Test | Tests that utilize the sums of squares for effects and the error. |

# The SAS code

```
/*Import data: student_mat from student performance dataset*/

proc import datafile="/home/u59571261/Final Project/student_mat.csv"

out=dat dbms=csv replace;

run; /*395 individuals with 34 variables*/
```

```
/*#################PROC TABULATE##################*/
/*PROC TABULATE: 1 variable*/
proc tabulate data=dat;
var G3;
class romantic;
tables romantic, G3*(N MEAN STD MIN MAX Q1 MEDIAN Q3); run;
/*PROC TABULATE: 2 variables*/
proc tabulate data=dat;
var G3;
class romantic famrel;
tables romantic, famrel*G3*(N MEAN); run;
/*PROC TABULATE: 3 variables*/
proc tabulate data=dat;
var G3;
class sex romantic famrel;
tables sex*romantic, famrel*G3*MEAN; run;

/*#################PROC SGPLOT#################*/
/*PROC SGPLOT: histogram/ density plot/ kornel density curve*/
proc sgplot data=dat;
histogram G3;
density G3;
density G3/ type=kernel; run;
/*PROC SGPLOT: horizontal box plot*/
proc sgplot data=dat;
hbox G3 / category=romantic; run;
/*PROC SGPLOT: vertical box plot*/
proc sgplot data=dat;
vbox G3 / category=famrel; run;

/*#################PROC GLM#################*/
/*PROC GLM: 1. Simple regression analysis*/
```

```
proc glm data=dat;
model G3=famrel; run; quit;
/*PROC GLM: 2. ANOVA analysis*/
proc glm data=dat;
class famrel;
model G3=famrel;
means famrel/ TUKEY; run; quit;
/*PROC GLM: 3. Two-way ANOVA analysis*/
proc glm data=dat;
class sex romantic;
model G3 = sex romantic sex*romantic/ ss3;
means sex romantic sex*romantic/bon; run; quit;
/*PROC GLM: 4. Repeated measures analysis of variance*/
proc glm data=dat;
class sex romantic;
model G1 G2 G3 = sex romantic sex*romantic/NOUNI;
repeated time 3 profile/ PRINTE ;
lsmeans sex romantic sex*romantic; run; quit;
```

## • Interpretation of the output from the SAS code

In this study, using the given data, 1) PROC TABULATE + PROC SGPLOT, and 2) PROC GLM SAS procedures were used for analysis.

The data used are as follows.

**Dataset**: Student_mat.csv file with 34 variables from student performance data

(https://archive.ics.uci.edu/ml/datasets/Student+Performance).

**Population**: A total of 395 individuals aged 15 to 22 years old.

**Purpose**: The purpose of this study was to determine the relationship between romantic relationship (yes or no), quality of family relationships (from 1 – very bad to 5 – excellent) and final grade (G3 variable) according to gender.

### PROC TABULATE + PROC SGPLOT

The following SAS program shows the number of subjects according to romantic relationship (Yes or No) and the final grade mean, standard deviation, minimum, maximum, Q1, median, and Q3 values using proc tabulate.

```
proc tabulate data=dat;

var G3;

class romantic;

tables romantic, G3*(N MEAN STD MIN MAX Q1 MEDIAN Q3); run;
```

As a result of <Table 1>, there were 263 cases of romantic relationship 'no' and 132 cases of 'yes'. The final grade (G3) mean of the group with a romantic relationship of ‘no’ was 10.84, the standard deviation was 4.39, the minimum value was 0, and the maximum value was 20. The final grade (G3) mean of the group with a romantic relationship of 'yes' was 9.58, the standard deviation was 4.86, the minimum value was 0, and the maximum value was 18.

The following SAS program shows the number of subjects and the average value of final grade according to romantic relationship and quality of family relationship using proc tabulate.

```
proc tabulate data=dat;

var G3;

class romantic famrel;

tables romantic, famrel*G3*(N MEAN); run;
```

As a result of <Table 2>, the mean of the final grade in the group with romantic relationship of 'no' and quality of family relationships of 1 was 6.67, in the group of 2, 9.85, in the group of 3, 10.47, in the group of 4, 10.98, 5 In the phosphorus group, it was confirmed that the average value was 11.17, which was high. On the other hand, when the romantic relationship is 'yes', it can be seen that the average of the final grade of the group with quality of family relationships of 1 is 13.00 the highest.

The following SAS program shows the number of subjects and the average value of final grade according to gender romantic relationship and quality of family relationship using proc tabulate.

```
proc tabulate data=dat;

var G3;

class sex romantic famrel;

tables sex*romantic, famrel*G3*MEAN; run;
```

As a result of <Table 3>, the mean of the final grade in the group with romantic relationship of 'no' and quality of family relationships of 1 was 6.67, in the group of 2, 9.85, in the group of 3, 10.47, and in the group of 4, 10.98, 5 In the phosphorus group, it was confirmed that the average value was 11.17, which was high. On the other hand, when the romantic relationship is 'yes', it can be seen that the average of the final grade of the group with quality of family relationships of 1 is 13.00 the highest.

The values displayed using the above proc tabulate syntax can be visualized using proc sgplot. The following shows the distribution of final grade scores in all subjects using proc sqplot using histogram, density plot, and kornel density curve <Figure 1>.

```
proc sgplot data=dat;

histogram G3;

density G3;

density G3/ type=kernel; run;
```

The following shows the mean and standard deviation of G3 according to romantic relationship using the proc sqplot syntax using the sgplot horizontal boxplot <Figure 2>.

```
proc sgplot data=dat;

hbox G3 / category=romantic; run;
```

The following shows the mean and standard deviation of G3 according to the quality of family relationship using the proc sqplot syntax using the sgplot vertical boxplot <Figure 3>.

```
proc sgplot data=dat;

vbox G3 / category=famrel; run;
```

## PROC GLM

### 1. Simple linear regression analysis

The following SAS program tries to verify the significant relationship between the two, using quality of family relationship as a continuous independent variable and final grade as a dependent variable. This analysis used simple linear regression analysis. In this case, 1) the independent variable (famrel) and the dependent variable (G3) should have a linear relationship, 2) the dependent variable should have a normal distribution at all values of the independent variable, 3) the residuals should be independent of each other. 4) It was analyzed assuming that the condition that the variance of the dependent variable would be the same for all independent variable values was satisfied.

```
proc glm data=dat;
```

```
model G3=famrel; run; quit;
```

Looking at the <Table 4> and <Figure 5> results, the final grade can be explained by this regression equation by 0.003% (R-square=0.003), and this regression equation was not statistically significant based on the significance level of 0.05. (F-value=1.04, p-value=0.309). When the quality of family relationship increased by 1, the final grade increased by 0.262 on average, which was not statistically significant (p-value=0.309). The estimated statistical formula is as follows.

## 2. Analysis of variance (ANOVA)

The following SAS program defines the quality of family relationship as a categorical independent variable, and attempts to compare the final grades in each independent group. Analysis of variance (ANOVA) was used for this analysis, and the size of each group was compared in two groups to see which group showed a significant difference through TUKEY's post-hoc analysis.

```
proc glm data=dat;

class famrel;

model G3=famrel;

means famrel/ TUKEY; run; quit;
```

Looking at the results of <Table 5> and <Figure 6>, the F-value was 0.40 and the p-value was 0,811, indicating that "the average final grade between the five groups is the same." the null hypothesis is not rejected. Looking at the results of TUKEY post-hoc analysis in <Table 6>, it can be confirmed that there is no group showing a statistically significant difference based on the significance level of 0.05. Therefore, the size of the final grade for each group of quality of family relationship does not differ.

## 3. Two-way ANOVA

Two-way ANOVA was performed to examine the main and interaction effects of gender and romantic relationship on the final grade. How the interaction effect appeared was confirmed through Bonferroni's multiple comparison.

```
proc glm data=dat;

class sex romantic;

model G3 = sex romantic sex*romantic;

means sex romantic sex*romantic/bon; run; quit;
```

Looking at <Table 7>, the main effect of gender on final grade was significant (F value=4.01, p-value <0.05), and the main effect of romantic status was also significant (F value=5.19, p- value <0.05). The interaction effect between gender and romantic status on final grade was not significant (F value=0.74, p-value= 0.392). <Figure 7> is a graphical representation of the interaction between gender and romantic status for the final grade. As a result of Bonferroni's multiple comparison, it was confirmed that there was a significant difference in final grade according to gender (p-value <0.05), and it was confirmed that there was a significant difference in final grade according to romantic status (p-value <0.05). 0.05) <Figure 8-9>. The mean and standard deviation of the final grade according to gender and romantic status can be confirmed in <Table 8>.

## 4. Repeated measure ANOVA

Mathematics scores were examined three times: first period grade (G1), second period grade (G2), and final grade (G3). Repeated measure ANOVA was used to analyze the repeated measure data. At this time, this analysis was performed assuming that 1) normality assumption, 2) sphericity assumption, and 3) the assumption of equality of covariance matrices were satisfied. In this analysis through repeated ANOVA, 1) within-subjects main effect to see whether the number of math credits for each group changes over time in gender and romantic relationship <Table 10-11>, 2) Differences in the number of math scores for each group in gender and romantic relationship The between-subjects main effects <Table 13> was checked to see whether there is

```
proc glm data=dat;
class sex romantic;
model G1 G2 G3 = sex romantic sex*romantic/NOUNI;
repeated time 3 profile/ PRINTE ;
lsmeans sex romantic sex*romantic; run; quit;
```

<Table 9> shows the results of Mauchly's sphericity tests. Since the Chi-square company is 95.78 and the p-value is less than 0.05, the null hypothesis that the data satisfies sphericity is rejected. Therefore, the modified test statistics (Greenhouse-Geisser correction (G-G) and the Huynh-Feldt correction (H-F)) are used. <Table 10> is the test result of the effect of time. The null hypothesis is that "there is no change in the mean grade over time". The null hypothesis is rejected with an F value of 8.72 and p <0.05. Therefore, there is a change in grade over time. <Table 11> is the interaction test result for each time, gender, and romantic status. First, in the case of gender, although the change in grade over time for females and males is the same (F value=1.13, p=0.323), the change in grade over time in the group with and without romantic relationship was statistically significant. Another was confirmed (F value=5.11, p-value <0.05). <Table 12> shows a hypothesis test of no romantic status by sex by time interaction. As a result of the analysis, the F value is 2.03 and p=0.133, so it cannot be said that there is a significant interaction between these variables. The results for the Between-subjects effect are as follows <Table 13>. There is no significant difference between men and women on their overall math grades (F value=3.54, p=0.061). The romantic status has no significant effect on overall math grades (F value=3.16, p=0.076). The interaction between sex and romantic status also shows a statistically nonsignificant result (F value=0.25, p=0.615). The univariate tests of hypotheses for within subjects effects were shown in Table 14. As mentioned above, since the sphericity assumption is not satisfied, the p-value of

G-G or H-F should be checked, not the F value. As a result, significant within-subjects main effects were confirmed with respect to time (p-value <0.05), and it was confirmed that there was a significant interaction effect between time and romantic status (p-value <0.05).