An Introduction to Proc SQL and Selected Statistical Methods

STAT 3505

Week 12 (April 04, 2024)

Gunes Fleming Ph.D.

TEMPLE UNIVERSITY

An Introduction to PROC SQL

An Introduction to PROC SQL

- PROC SQL is a powerful Base SAS Procedure that combines the functionality of DATA and PROC steps into a single step.

- PROC SQL can sort, summarize, subset, join (merge), and concatenate datasets, create new variables, and print the results or create a new table or view all in one step!

- PROC SQL can not only retrieve information without having to learn SAS syntax, but it can often do this with fewer and shorter statements than traditional SAS code.

- Additionally, SQL often uses fewer resources than conventional DATA and PROC steps. Further, the knowledge learned is transferable to other SQL packages.

An Introduction to PROC SQL

- In general, the Structured Query Language (SQL) is a standardized language used to retrieve and update data stored in relational tables (or databases).

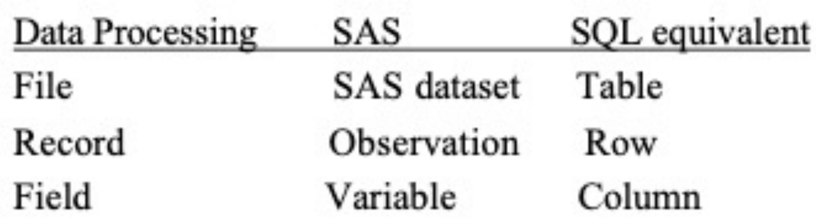
- When coding in SQL, the user is not required to know the physical attributes of the table such as data location and type. SQL is non- procedural. The purpose is to allow the programmer to focus on what data should be selected and not how to select the data. The method of retrieval is determined by the SQL optimizer, not by the user.

- A table is a two-dimensional representation of data consisting of columns and rows. A table in SQL is simply another term for a SAS data set.

- Tables are logically related by values such as a key column.

An Introduction to PROC SQL

- Terminology:



Data Processing	SAS	SQL equivalent
File	SAS dataset	Table
Record	Observation	Row
Field	Variable	Column

- The table is where the data is stored. A row represents a particular entry. An employee may be represented as a row in a table. A column represents the particular values for all rows. Salary may be a column on a table. All employees will have a value for salary.

An Introduction to PROC SQL

- Simple Queries: A query is a request for information from a table or tables. The query result is typically a report but can also be another table.

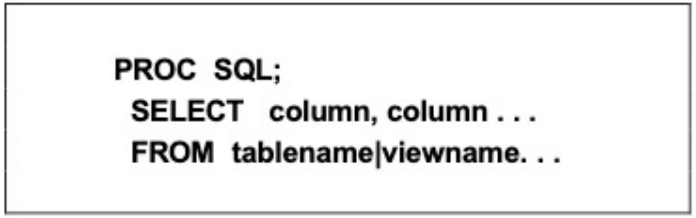
- For instance: I would like to select last name, department, and salary from the employee table where the employee's salary is greater than 35,000.

How would this query (request) look in SQL?

```
SELECT LASTNAME, DEPARTMENT, SALARY
FROM CLASS.EMPLOY
WHERE SALARY GT 35000
```

An Introduction to PROC SQL

- Basic Proc SQL Syntax:

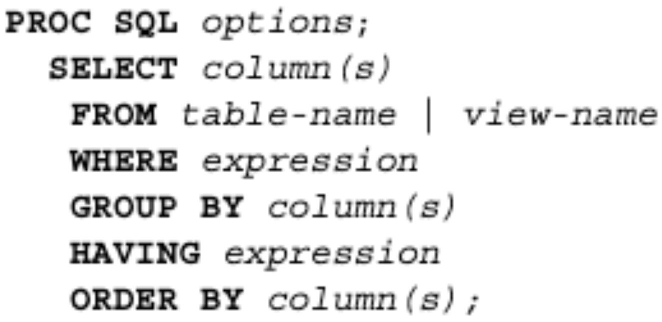


```
PROC  SQL;
 SELECT   column, column . . .
 FROM  tablename|viewname. . .
```

- In Proc SQL,
  - Statements (clauses) in the SQL procedure are not separated by semicolons, the entire query is terminated with a semicolon.
  - Items in an SQL statement are separated by a comma.
  - There is a required order of statements in a query.
  - One SQL procedure can contain many queries and a query can reference the results from previous queries.
  - The SQL procedure can be terminated with a QUIT statement, RUN statements have no effect.

An Introduction to **PROC SQL**

- Select Statement:
  - To retrieve and display data a SELECT statement is used.
  - The data will be displayed in the order you list the columns in the SELECT statement.
  - A column can be a variable, calculated value, assigned value or formatted value.
  - An asterisk (*) can be used to select all columns.
- SELECT Syntax:



```
PROC SQL options;
  SELECT column(s)
   FROM table-name | view-name
   WHERE expression
   GROUP BY column(s)
   HAVING expression
   ORDER BY column(s);
```

An Introduction to **PROC SQL**

- *From* (Part of Select Statement): Specifies the input table(s).

- For example, to select social security number, salary and bonus (columns) for all employees (rows) from the employeedata table (data set):

```
PROC SQL;
          SELECT SSN, SALARY, BONUS
          FROM CLASS.EMPLOYEEDATA;
QUIT;
```

PROC SQL Basics

- An asterisk can be used on the SELECT statement to select all columns.
- To specify certain variables, variables can be listed on the SELECT statement.
- The AS keyword can be used to rename variables when selecting or creating new variables in the output.
- To refer to the results of calculations on columns within a select statement, the keyword CALCULATED must be used.

PROC SQL Basics

- The CASE statement is used to perform conditional logic within a SELECT statement.
- CASE statement allows you to evaluate conditions and return different values based on those conditions. It is useful for creating derived variables or transforming existing variables based on specific criteria.
- END is required when using the CASE.
- Coding the WHEN in descending order of probability will improve efficiency because SAS will stop checking the CASE conditions as soon as it finds the first true value.

PROC SQL Basics

- Summary functions such as AVG/MEAN, COUNT/FREQ, MAX, MIN, NMISS, STD, SUM, VAR can be used on the SELECT Statement.
- Unless GROUP BY is present when a summary function is used, the summary statistics would be remerged to the dataset.
- GROUP BY clause is used to group rows that have the same values into summary rows, often to perform aggregate summaries on the grouped data.
- It also sorts the data by the grouping variables.

PROC SQL Basics

- The ORDER BY clause will return the data in sorted order. In the absence of summary functions, ORDER BY can be used to sort data.
- In order to subset data, WHERE can be used similar to a DATA step.
- As part of WHERE, a calculated variable cannot be directly referred. Either the calculation can be repeated in WHERE clause, or the keyword CALCULATED can be used along with the derived column name.
- When grouping is in affect (GROUP BY), WHERE cannot be used for subsetting. Instead, HAVING can be used.

PROC SQL Basics

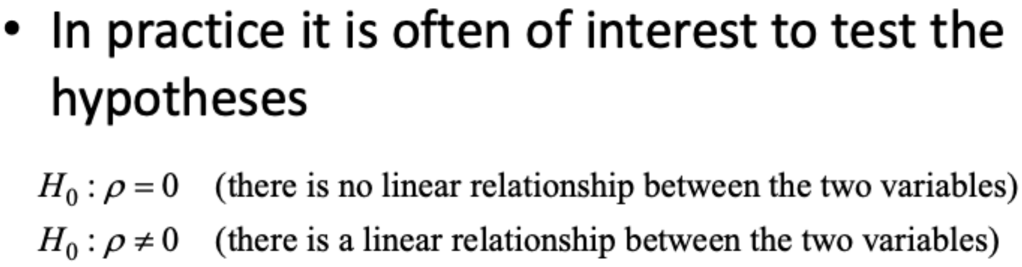
- To create a table, “CREATE TABLE *tablename* AS” must be used right before the SELECT statement.
- To concatenate two datasets, UNION operator can be used. The UNION operator keeps only unique observations. To keep all observations, the UNION ALL operator can be used.
- To merge two or more datasets, INNER/LEFT/RIGHT JOIN can be used.

Correlation and Regression

- Correlation measures the association between two quantitative variables.
- Simple linear regression creates an equation to predict the value of the dependent variable using a single independent variable.
- Multiple linear regression creates an equation to predict the value of the dependent variable using multiple independent variables.

Correlation

- The correlation coefficient is a measure of the linear relationship between two quantitative variables measured on the same subject (or entity).
- The sample correlation r is a unitless quantity (i.e. it does not depend on the units of measurement) that ranges from –1 to +1.
- The correlation coefficient rho is typically estimated from data using the Pearson correlation coefficient and designated as r.



- In practice it is often of interest to test the hypotheses

$H_0 : \rho = 0$ (there is no linear relationship between the two variables)

$H_0 : \rho \neq 0$ (there is a linear relationship between the two variables)

Correlation

- In practice it is often of interest to test the hypotheses

$H_0 : \rho = 0$ (there is no linear relationship between the two variables)

$H_0 : \rho \neq 0$ (there is a linear relationship between the two variables)

- PROC CORR in SAS provides a test of the above hypotheses designed to determine whether the estimated correlation coefficient, r, is significantly different from zero.
- This test assumes that the data represent a random sample from some bivariate normal population. If normality is not a good assumption, nonparametric correlation estimates are available, the most popular of which is Spearman’s rho, which PROC CORR also provides.
- To examine the nature of the relationship between two variables is always good practice to look at scatterplots of the variables.

Correlation

Syntax

PROC CORR <options>; <statements>;

- Common Options:
  - **DATA=datsetname**; Specifies dataset.
  - **SPEARMAN** - Requests Spearman rank correlations.
  - **NOSIMPLE -** Suppresses display of descriptive
  - **NOPROB -** Suppresses display of p-values
  - **PLOTS** – Produces plots

Common Statements

- **VAR variable(s);** Pair wise correlations are calculated for the variables listed.
- **BY variable(s);** Produces separate set of pairwise correlations for variables in the VAR list for each level of the categorical variable in the BY list. (Data must be sorted prior to using the BY statement in PROC CORR.)
- **WITH variable(s);** Correlations are obtained between the variables in the VAR list with variables in the WITH list

Simple Linear Regression

- The regression line that SAS calculates from the data is an estimate of a theoretical line describing the relationship between the independent variable (X) and the dependent variable (Y).
- The theoretical line is $Y = \alpha + \beta x + \varepsilon$

where α is the y-intercept, β is the slope, and ε is an error term that is normally distributed with zero mean and constant variance.

- It should be noted that β = 0 indicates that there is no linear relationship between X and Y.
- A simple linear regression analysis is used to develop an equation (a linear regression line) for predicting the dependent variable given a value (x) of the independent variable.

Simple Linear Regression

- The regression line calculated by SAS is given by $\hat{Y} = a + bx$

  where a and b are the least-squares estimates of α and β.

- The null hypothesis that there is no predictive linear relationship between the two variables is that the slope of the regression equation is zero.

- Specifically, the hypotheses are: $$H_0: \beta = 0$$ $$H_a: \beta \neq 0$$

- A low p-value for this test (say, less than 0.05) indicates significant evidence to conclude that the slope of the line is not 0; that is, that knowledge of X would be useful in predicting Y.

Simple Linear Regression

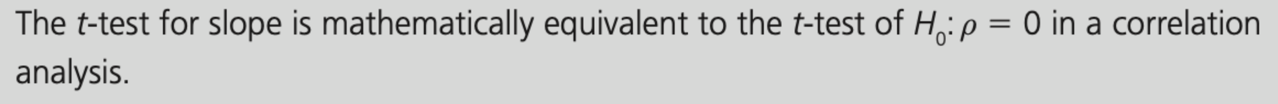
- The regression line calculated by SAS is given by $\hat{Y} = a + bx$

  where a and b are the least-squares estimates of α and β.

- The null hypothesis that there is no predictive linear relationship between the two variables is that the slope of the regression equation is zero.

- Specifically, the hypotheses are: $H_0\!: \beta = 0$ $H_a\!: \beta \neq 0$

- A low p-value for this test (say, less than 0.05) indicates significant evidence to conclude that the slope of the line is not 0; that is, that knowledge of X would be useful in predicting Y.



The *t*-test for slope is mathematically equivalent to the *t*-test of $H_0\!: \rho = 0$ in a correlation analysis.

Simple Linear Regression: PROC REG

Syntax

- The general syntax for PROC REG is

PROC REG <options>; <statements>;

- Common OPTIONS are
  - **DATA=*datsetname* -** Specifies dataset.
  - **SIMPLE -** Displays descriptive statistics.

Common Statements

- **MODEL** dependentvar = independentvar</options>;
- **BY groupvariable;** Produces separate regression analyses for each value of the BY variable

- Note: Several MODEL options are available, but we will defer their discussion to the section on using SAS for multiple linear regression.

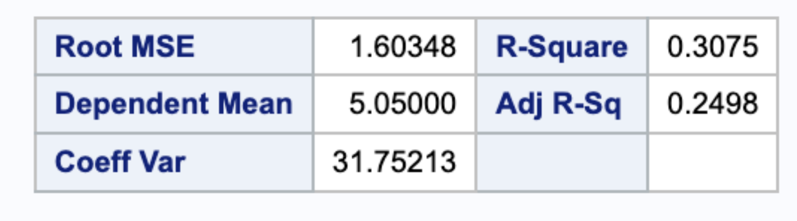
Simple Linear Regression: PROC REG

- Example: A random sample of fourteen elementary school students is selected from a school, and each student is measured on a creativity score (X) using a new testing instrument and on a task score (Y) using a standard instrument. The task score is the mean time taken to perform several hand–eye coordination tasks. Because administering the creativity test is much cheaper, the researcher wants to know if the CREATE score is a good substitute for the more expensive TASK score.

Simple Linear Regression: PROC REG

- Example (Continues):

SAS Results

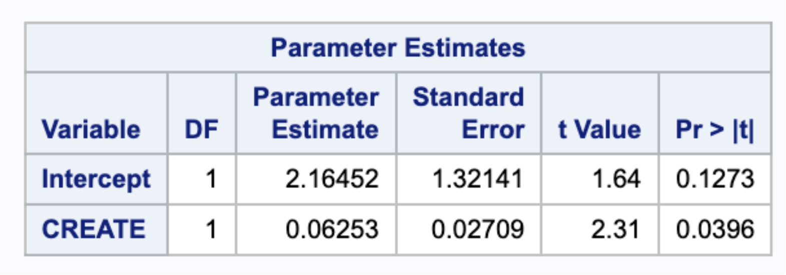


Root MSE	1.60348	**R-Square**	0.3075
Dependent Mean	5.05000	**Adj R-Sq**	0.2498
Coeff Var	31.75213		

- R-Square: This is a measure of the strength of the association between the dependent and independent variables. (It is the square of the Pearson correlation coefficient.) The closer this value is to 1, the stronger the association. In this case R2 = 0.31 indicates that 31% of the variability in TASK is explained by the regression with CREATE.

Simple Linear Regression: PROC REG

- Example (Continues): <u>SAS Results</u>

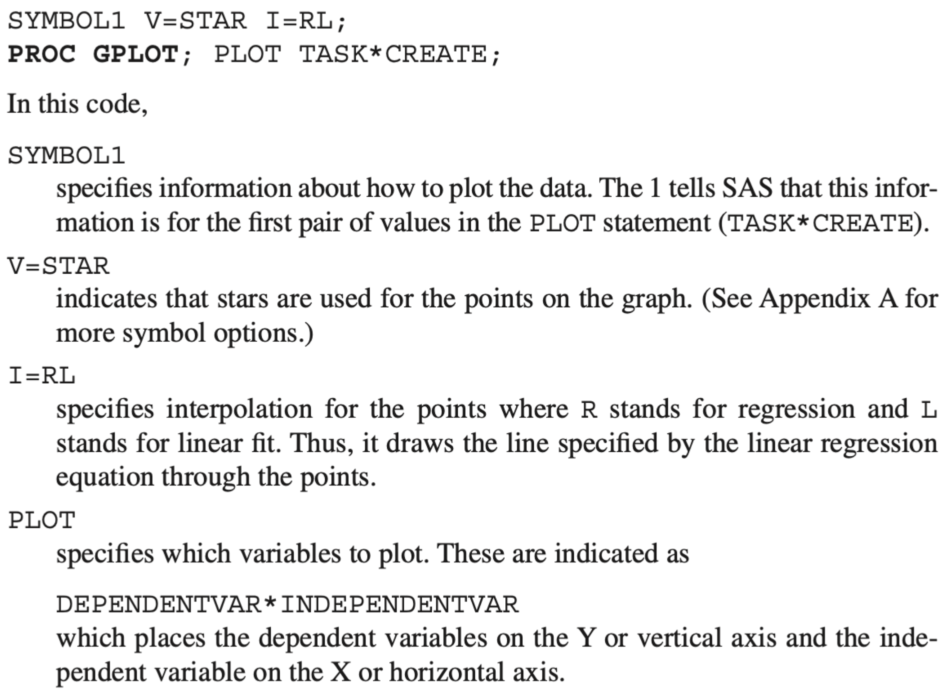


Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.16452	1.32141	1.64	0.1273
CREATE	1	0.06253	0.02709	2.31	0.0396

- Slope: The statistical test on the “CREATE” row is for a test of H0: $\beta = 0$, and $p = 0.04$ provides evidence to reject the null hypothesis and conclude that the slope is not zero. (The statistical test on the Intercept is generally of little importance.)
- Estimates: The column labeled “Parameter Estimate” gives the least squares estimates a and b of the regression equation. In this case, the equation is: TASK = 2.16452 + 0.06235 * CREATE;
- Thus, from this equation you can predict a value of the TASK score from the CREATE score. However, before making any predictions using this equation, you should analyze the relationship further.

Creating a Simple Linear Regression Plot:

- It is always a good idea to plot your data to examine the linearity of the relationship. In SAS, you can create a simple scatterplot by adding the following statements to the code:



```
SYMBOL1 V=STAR I=RL;
PROC GPLOT; PLOT TASK*CREATE;
```

In this code,

SYMBOL1

specifies information about how to plot the data. The 1 tells SAS that this information is for the first pair of values in the PLOT statement (TASK*CREATE).

V=STAR

indicates that stars are used for the points on the graph. (See Appendix A for more symbol options.)

I=RL

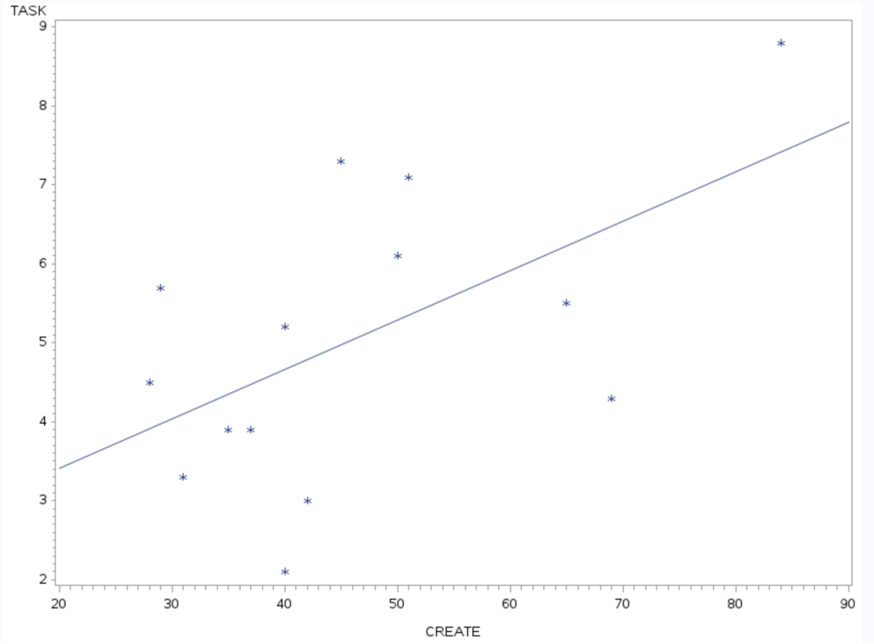
specifies interpolation for the points where R stands for regression and L stands for linear fit. Thus, it draws the line specified by the linear regression equation through the points.

PLOT

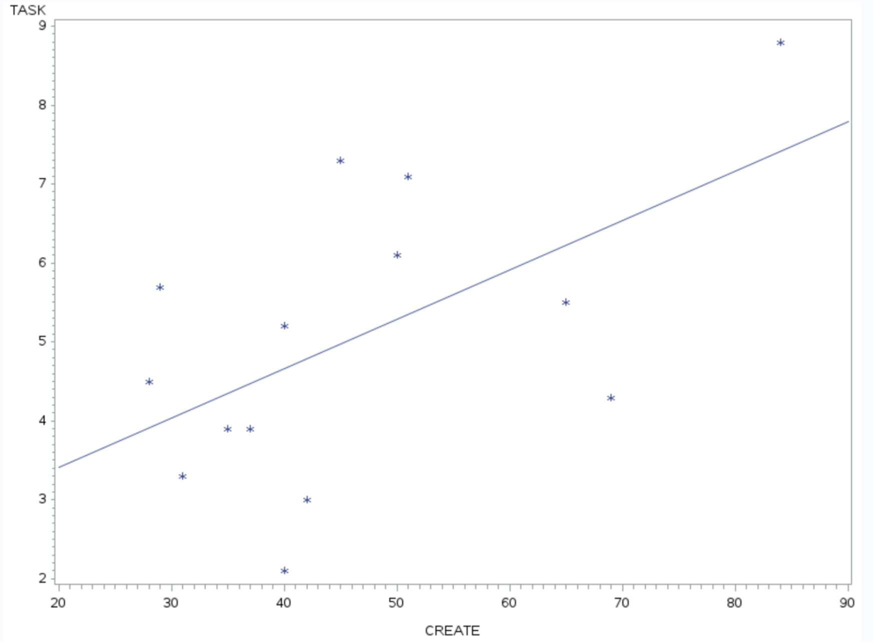
specifies which variables to plot. These are indicated as

DEPENDENTVAR*INDEPENDENTVAR

which places the dependent variables on the Y or vertical axis and the independent variable on the X or horizontal axis.



Creating a Simple Linear Regression Plot:



- In the plot, the independent variable (CREATE) is on the x-axis and the dependent variable (TASK) is on the y-axis.

- By observing the scatterplot, a positive correlation between the two variables can be seen (in this case r = 0.74), and it appears that knowing CREATE should help in predicting TASK.

- It is also clear that knowing CREATE does not in any way perfectly predict TASK.