#SASGF

VIRTUAL

SAS® GLOBAL FORUM 2021

AMERICAS | MAY 18 – 20 ASIA PACIFIC | MAY 19 – 20 EMEA | MAY 25 – 26

Reporting Correlation Coefficient results and Plots – A SAS® Macro that does it all

Zhengming Chen, Weill Cornell Medicine

My Brief bio

I am an Assistant Professor of Biostatistics from Weill Cornell Medicine in New York. I provide statistical support and do biomedical research for the medical school. I also teach a course *Statistical Programming with SAS* for our *Biostatistics and Data Science* graduate program

Outlines

- Background
- SAS implementation of correlation analyses
- Build a SAS Macro to report different types of correlation with table and figure
  - Strategy and workflow
  - The **`%CorrReport`** Macro
  - Demo how to use the Macro
  - Limitations

Background

Motivation

- Table-1 in medical research
  - Summary statistics:
    - Numerical variables: N, Mean(SD), Median(IQR), missing, etc.
    - Categorical variables: N, Proportion, missing
  - Bivariate association:
    - **Table-1 is often a two-way cross table: all variables stratified by a Categorical variable**
    - **Significance test and strength of association:**

      1). Numerical vs. Categorical: *t*-test, ANOVA; nonparametric test; trend test;

      2). Categorical vs. Categorical: Chi-squared test (Fisher's exact test); trend test; agreement test;

Background

Motivation

- Examples of Table-1 in journals

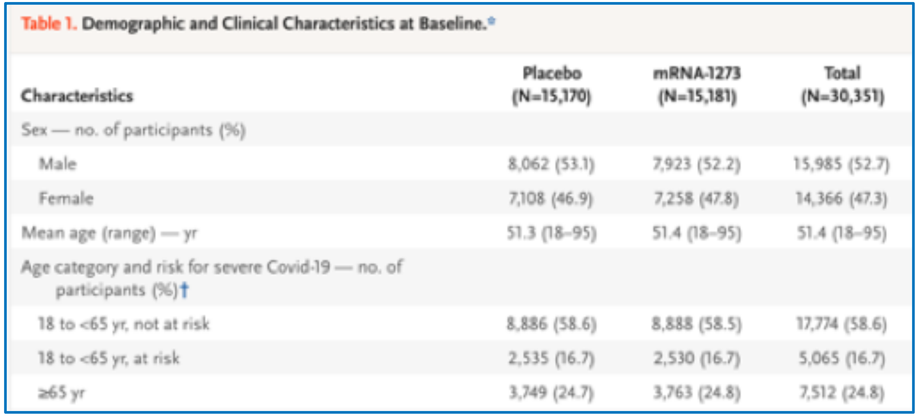


Table 1. **Demographic and Clinical Characteristics at Baseline.***

Characteristics	Placebo (N=15,170)	mRNA-1273 (N=15,181)	Total (N=30,351)
Sex — no. of participants (%)			
Male	8,062 (53.1)	7,923 (52.2)	15,985 (52.7)
Female	7,108 (46.9)	7,258 (47.8)	14,366 (47.3)
Mean age (range) — yr	51.3 (18–95)	51.4 (18–95)	51.4 (18–95)
Age category and risk for severe Covid-19 — no. of participants (%)†			
18 to <65 yr, not at risk	8,886 (58.6)	8,888 (58.5)	17,774 (58.6)
18 to <65 yr, at risk	2,535 (16.7)	2,530 (16.7)	5,065 (16.7)
≥65 yr	3,749 (24.7)	3,763 (24.8)	7,512 (24.8)

N Engl J Med; 2021 Feb 4;384(5):403-416.

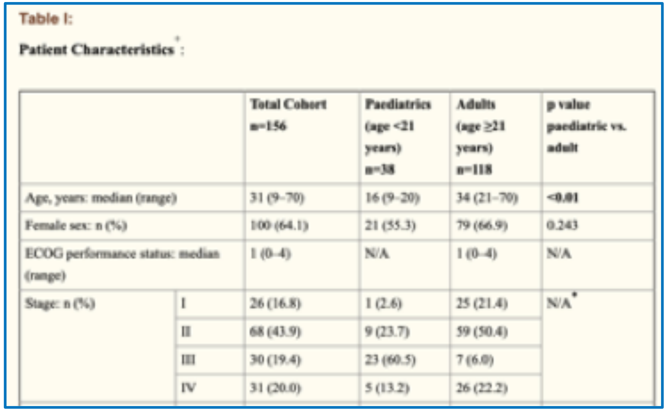


Table I:

Patient Characteristics*:

		Total Cohort n=156	Paediatrics (age <21 years) n=38	Adults (age ≥21 years) n=118	p value paediatric vs. adult
Age, years: median (range)		31 (9–70)	16 (9–20)	34 (21–70)	<0.01
Female sex: n (%)		100 (64.1)	21 (55.3)	79 (66.9)	0.243
ECOG performance status: median (range)		1 (0–4)	N/A	1 (0–4)	N/A
Stage: n (%)	I	26 (16.8)	1 (2.6)	25 (21.4)	N/A*
	II	68 (43.9)	9 (23.7)	59 (50.4)	
	III	30 (19.4)	23 (60.5)	7 (6.0)	
	IV	31 (20.0)	5 (13.2)	26 (22.2)	

Br J Haematol; 2017 Dec;179(5):739-747.

#SASGF

SAS GLOBAL FORUM 2021

Background

Motivation

- SAS Macro to produce Table-1
  - SAS procedures to produce the results for Table-1:

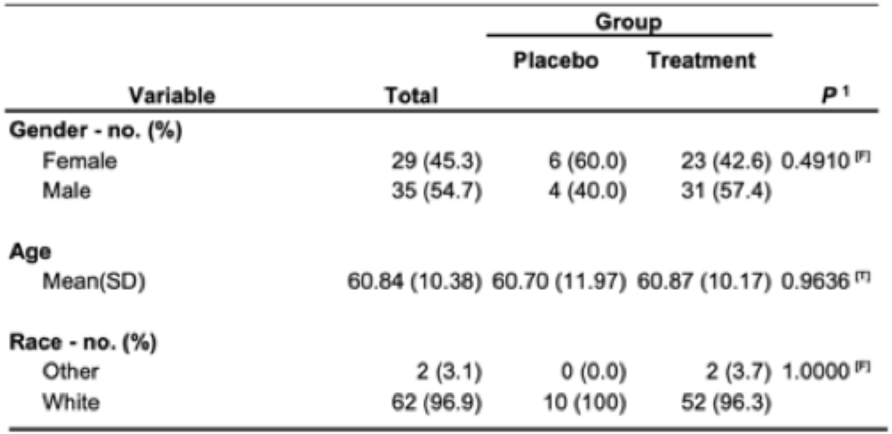
    `Proc MEANS; Proc FREQ; Proc TTEST; Proc GLM; Proc NPAR1WAY;` etc.
  - SAS Macro
    - Run the Procs -> output the results -> combine the outputs -> report with ODS
    - Reproducible, efficient and productive
    - Examples:

      `%Table1Macro; %Table1nDone; %SummaryTable; %Table_summary; %table1; %ggBaseline,` etc.

Background

Motivation

- SAS Macro to produce Table-1
  - Example output from a SAS Macro



		Group		
Variable	**Total**	**Placebo**	**Treatment**	***P*** [1]
Gender - no. (%)				
Female	29 (45.3)	6 (60.0)	23 (42.6)	0.4910 [F]
Male	35 (54.7)	4 (40.0)	31 (57.4)	
Age				
Mean(SD)	60.84 (10.38)	60.70 (11.97)	60.87 (10.17)	0.9636 [T]
Race - no. (%)				
Other	2 (3.1)	0 (0.0)	2 (3.7)	1.0000 [F]
White	62 (96.9)	10 (100)	52 (96.3)	

Background

Motivation

- One type of bivariate relationship is missing in Table-1
  - Numerical vs. numericcal variable
    - Summary statistics and correlation coefficient
  - No dedicated SAS Macro for general correlation analysis and reporting like the ones for Table-1
    - Some specialized Macro for specific types of coefficients:
      - ➢ Intraclass correlation coefficients (**`%icc9`**)
      - ➢ Compute biserial, point biserial, and rank biserial correlations between a binary and a continuous (or ranked) variable (**`%BISERIAL`**)

Background

Motivation

- A SAS Macro for correlation analysis to supplement Table-1 is needed
  - For practical use: reproducible and productive
  - A teaching example:
    - ➢ How to build a SAS Macro from scratch: data step, Proc, ODS, figures, Macro, etc.
    - ➢ A complete cycle of a statistical analysis: prepare data, analyze, report in Table and Figures, etc.

Background

Correlation Coefficient

- Pearson Correlation Coefficient
  - *A descriptive measure of the degree and direction of linear relationship between two continuous variables when they are random variables and follow bivariate normal distribution*
  - Math:
    - Population

$$\rho = Corr(X, Y) = Cov(X, Y)/(\sigma_x \sigma_y) = E[(X - \mu_x)(Y - \mu_y)]/(\sigma_x \sigma_y) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \mu_x}{\sigma_x} \right) \left(\frac{y_i - \mu_y}{\sigma_y} \right)$$

    - Sample

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Or: $r_{xy} = \pm\sqrt{R^2}$

  - Example

Patient’s height vs. weight

Background

Correlation Coefficient

- Spearman Rank-Correlation Coefficient
  - *A nonparametric measure of correlation based on ranks of the data values*
  - Math:

$$\theta = \frac{\sum_i ((R_i - \bar{R})(S_i - \bar{S}))}{\sqrt{\sum_i (R_i - \bar{R})^2 \sum (S_i - \bar{S})^2}}$$

where R_i is the rank of x_i , S_i is the rank of y_i , $\bar{R}$ is the mean of the R_i values, and $\bar{S}$ is the mean of the S_i values.

  - Example:

Patient's survival time after treatment vs. Age

Background

Correlation Coefficient

- Polychoric Correlation
  - *Correlation between two unobserved continuous variables that have a bivariate normal distribution. The unobserved information is obtained from two observed ordinal variables.*
  - Math:

    The polychoric correlation coefficient is the **maximum likelihood estimate** of the product-moment correlation between the underlying normal variables.
  - Example:

    Patient's quality of life scale (1 - 10) vs. Severity of Covid-19 symptom (1 - 5)

Background

Correlation Coefficient

- Polyserial Correlation
  - *Correlation between two continuous variables that have a bivariate normal distribution, where one variable is observed directly, and the other is unobserved but an ordinal variable.*
  - **Math:**

    By **maximum likelihood estimate** of a set of parameters
  - Example:

    Patient's BMI vs. Patient's satisfaction scale (1 – 10)

Background

Correlation Coefficient

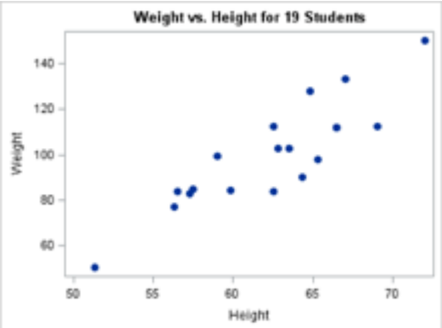
- Correlation Coefficients are always between -1 and 1, the correlation is stronger when it is more away from 0.
- The sign of Correlation Coefficient shows the direction of the correlation.

Background

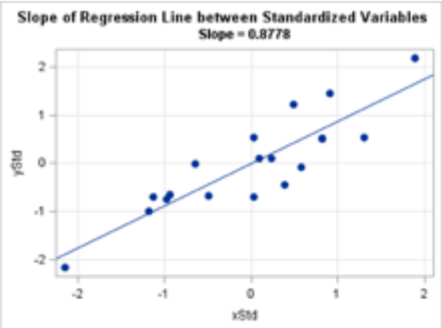
Correlation Coefficient

- To visualize the correlations
  - Thirteen Ways to Look at the Correlation Coefficient (by *Joseph Lee Rodgers* and *W. Alan Nicewander*, 1988)
  - 7 Ways to view correlation (by *Rick Wicklin*, 2017)
    - https://blogs.sas.com/content/iml/2017/09/05/7-ways-view-correlation.html
    - Graphically:

Scatter plot



Slope of regression line of standardized data



SAS implementation of correlation analyses

Apply Fisher’s z-transformation to obtain the confidence interval of ***r***

“Pearson” is the default

```
Proc CORR data=sashelp.class Pearson Spearman
       fisher(rho0 = 0) Polychoric Polyserial
       plots=all;
 var height;
 with weight;
run;
```

To set the null ***r*** for hypothesis testing. Default is 0

To request different types of correlation depending on the type of variables

“plots” is to show plots

Output:

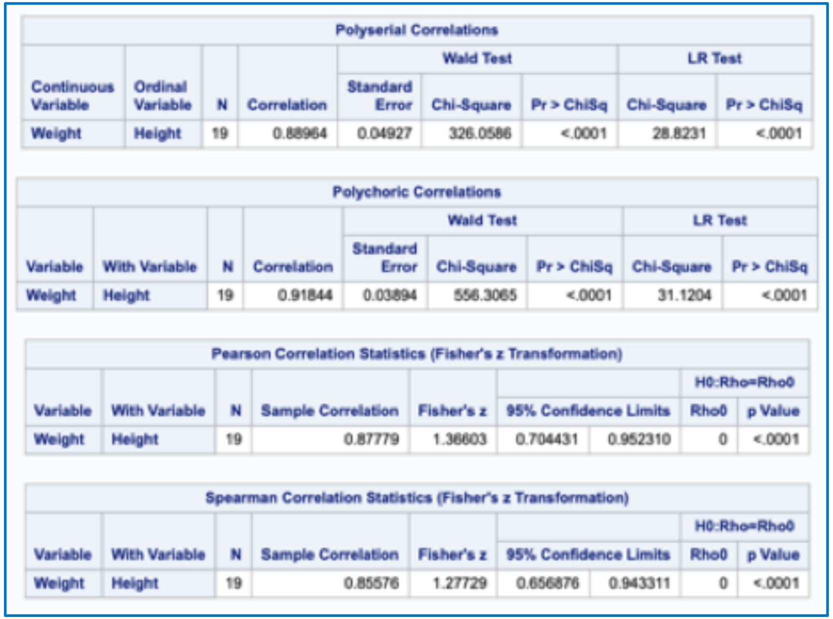
The CORR Procedure

1 With Variables:	Height
1 Variables:	Weight

Simple Statistics						
Variable	**N**	**Mean**	**Std Dev**	**Median**	**Minimum**	**Maximum**
Height	19	62.33684	5.12708	62.80000	51.30000	72.00000
Weight	19	100.02632	22.77393	99.50000	50.50000	150.00000

Pearson Correlation Coefficients, N = 19 Prob > |r| under H0: Rho=0	
	Weight
Height	0.87779 <.0001

Spearman Correlation Coefficients, N = 19 Prob > |r| under H0: Rho=0	
	Weight
Height	0.85576 <.0001



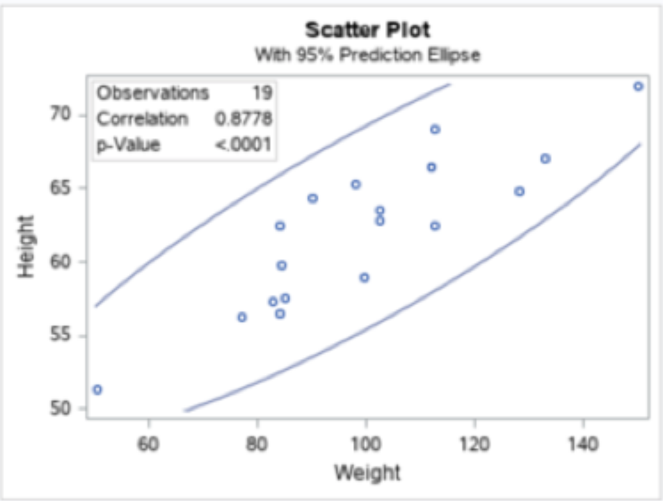
Polyserial Correlations								
				Wald Test			**LR Test**	
Continuous Variable	**Ordinal Variable**	**N**	**Correlation**	**Standard Error**	**Chi-Square**	**Pr > ChiSq**	**Chi-Square**	**Pr > ChiSq**
Weight	**Height**	19	0.88964	0.04927	326.0586	<.0001	28.8231	<.0001

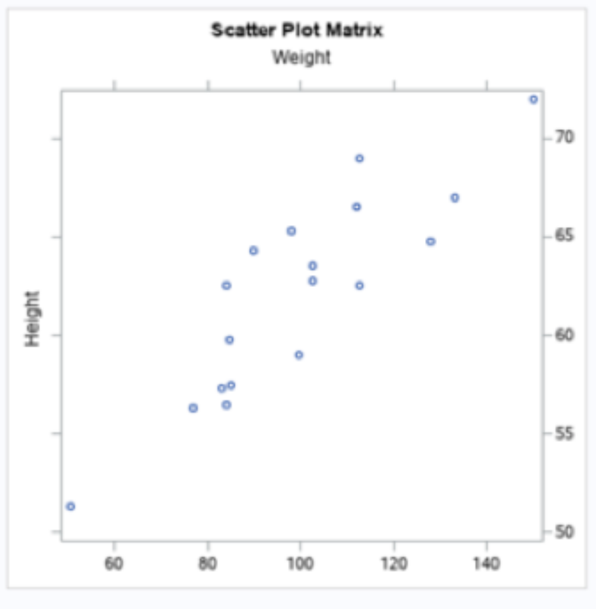
Polychoric Correlations								
				Wald Test			**LR Test**	
Variable	**With Variable**	**N**	**Correlation**	**Standard Error**	**Chi-Square**	**Pr > ChiSq**	**Chi-Square**	**Pr > ChiSq**
Weight	**Height**	19	0.91844	0.03894	556.3065	<.0001	31.1204	<.0001

Pearson Correlation Statistics (Fisher's z Transformation)								
							H0:Rho=Rho0	
Variable	**With Variable**	**N**	**Sample Correlation**	**Fisher's z**	**95% Confidence Limits**		**Rho0**	**p Value**
Weight	**Height**	19	0.87779	1.36603	0.704431	0.952310	0	<.0001

Spearman Correlation Statistics (Fisher's z Transformation)								
							H0:Rho=Rho0	
Variable	**With Variable**	**N**	**Sample Correlation**	**Fisher's z**	**95% Confidence Limits**		**Rho0**	**p Value**
Weight	**Height**	19	0.85576	1.27729	0.656876	0.943311	0	<.0001

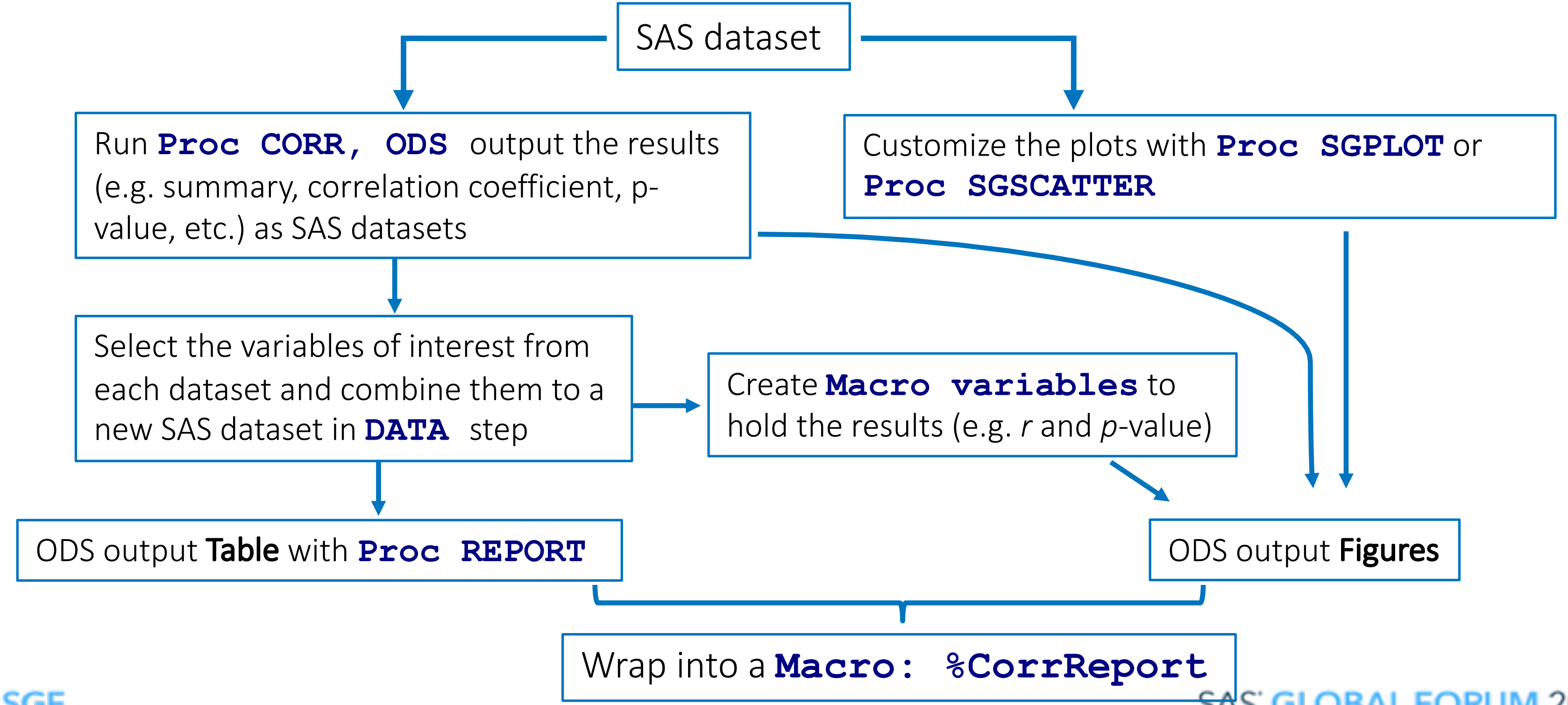
Plots:


Scatter Plot
With 95% Prediction Ellipse
Observations 19
Correlation 0.8778
p-Value <.0001
Height
70
65
60
55
50
60
80
100
120
140
Weight


Scatter Plot Matrix
Weight
Height
70
65
60
55
50
60
80
100
120
140

Build a SAS Macro to report correlation analyses

Strategy and workflow



SAS dataset

Run **`Proc CORR, ODS`** output the results (e.g. summary, correlation coefficient, p-value, etc.) as SAS datasets

Customize the plots with **`Proc SGPLOT`** or **`Proc SGSCATTER`**

Select the variables of interest from each dataset and combine them to a new SAS dataset in **`DATA`** step

Create **`Macro variables`** to hold the results (e.g. *r* and *p*-value)

ODS output **Table** with **`Proc REPORT`**

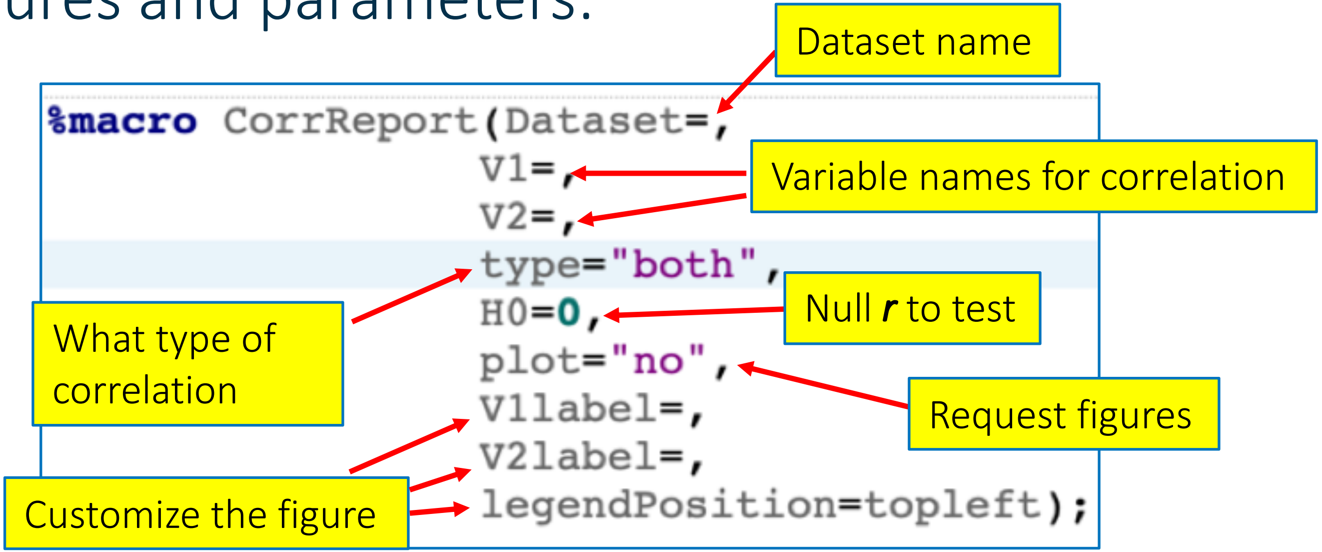
ODS output **Figures**

Wrap into a **`Macro: %CorrReport`**

Build a SAS Macro to report correlation analyses

The %CorrReport Macro

- Features and parameters:



```
%macro CorrReport(Dataset=,
                  V1=,
                  V2=,
                  type="both",
                  H0=0,
                  plot="no",
                  V1label=,
                  V2label=,
                  legendPosition=topleft);
```

Dataset name

Variable names for correlation

What type of correlation

Null *r* to test

Request figures

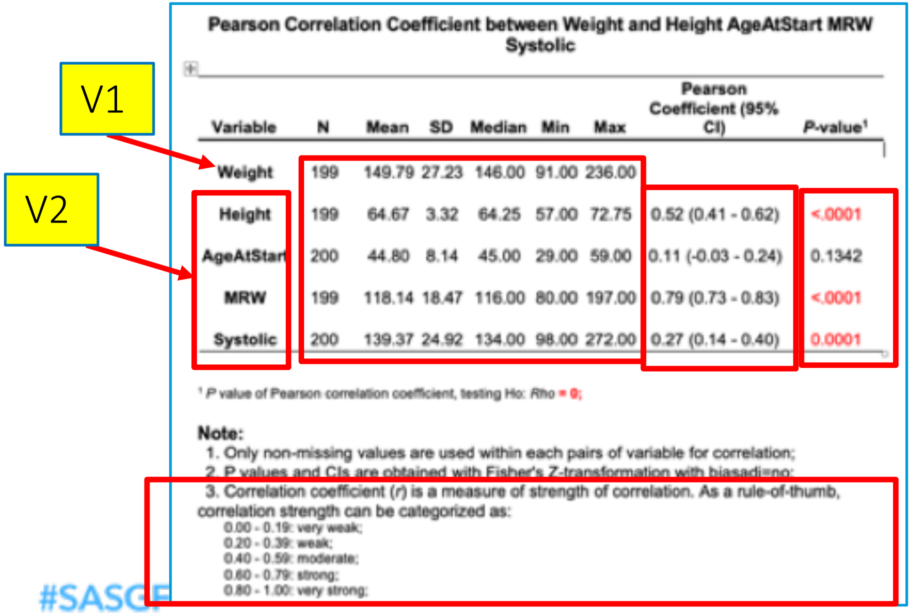
Customize the figure

Build a SAS Macro to report correlation analyses

The %CorrReport Macro

- Demo how to use the Macro (in SAS Studio)

Table:



V1

V2

Pearson Correlation Coefficient between Weight and Height AgeAtStart MRW Systolic

Variable	N	Mean	SD	Median	Min	Max	Pearson Coefficient (95% CI)	*P*-value[1]
Weight	199	149.79	27.23	146.00	91.00	236.00		
Height	199	64.67	3.32	64.25	57.00	72.75	0.52 (0.41 - 0.62)	<.0001
AgeAtStart	200	44.80	8.14	45.00	29.00	59.00	0.11 (-0.03 - 0.24)	0.1342
MRW	199	118.14	18.47	116.00	80.00	197.00	0.79 (0.73 - 0.83)	<.0001
Systolic	200	139.37	24.92	134.00	98.00	272.00	0.27 (0.14 - 0.40)	0.0001

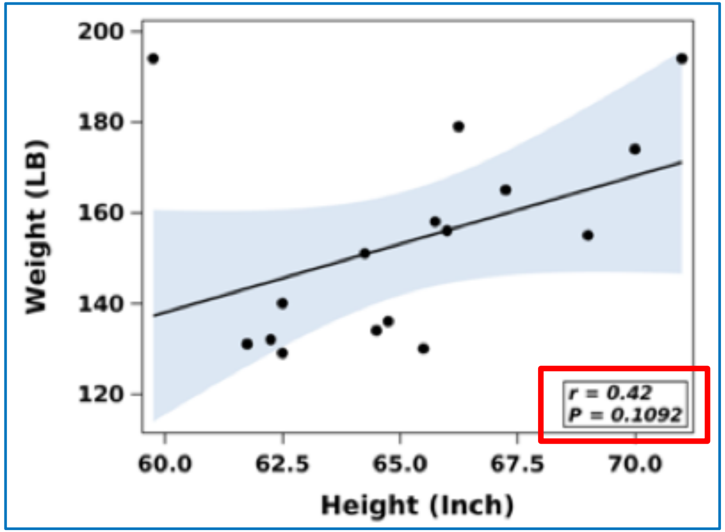
[1] *P* value of Pearson correlation coefficient, testing Ho: *Rho* = 0;

Note:

1. Only non-missing values are used within each pairs of variable for correlation;
2. P values and CIs are obtained with Fisher's Z-transformation with biasadj=no;
3. Correlation coefficient (*r*) is a measure of strength of correlation. As a rule-of-thumb, correlation strength can be categorized as:

0.00 - 0.19: very weak;
0.20 - 0.39: weak;
0.40 - 0.59: moderate;
0.60 - 0.79: strong;
0.80 - 1.00: very strong;

Figure:



Build a SAS Macro to report correlation analyses

The %CorrReport Macro

- Limitations and improvement
    - No customized error messages yet
    - Kendall’s Tau-b Correlation Coefficient
    - Pearson, Spearman, and Kendall partial correlation
    - Cronbach’s Coefficient Alpha

Takeaways

- SAS is comprehensive in correlation analyses
- SAS Macro is powerful for reproducible and efficient analysis and reporting
- This SAS Macro is useful tool in real world practice and in class room. It covers a complete cycle of data analysis with SAS.
- The skills in building this Macro are extendable to other Macro…

Thank you!

Contact Information
zhc2006@med.cornell.edu