Missing Data with Correlation & Multiple Regression

Missing Data

Missing data have several sources, response refusal, coding error, data entry errors, and outliers are a few. SPSS allows you to identify specific data values as “missing” – those specific values will be recognized as “non data” and not used in statistical computations. Once the missing values are set, it is easy to use Frequencies to find the number of cases with missing data for each variable

This data set of N = 103 cases has no more than 6 missing values for any variable – so, around 1-5% outliers, not bad.

But remember that we are using at least 2 (a single correlation) and maybe many more (several correlations or a multiple regression) variables in our analyses.

The real problem with missing data is that the number of cases with incomplete data “adds up” across the multiple variables used in an analysis

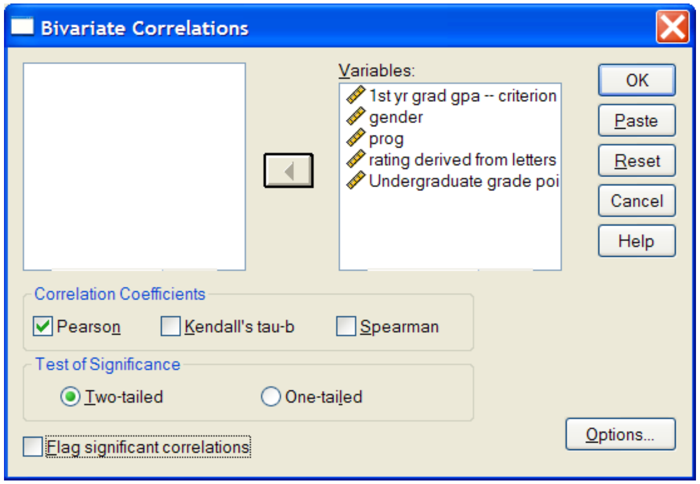
“Gender” is coded 2 = in Gender Studies Concentration 1 = not
“Prog” is coded 2 = in Clinical Program 1 = in Experimental Program

Statistics

		1st yr grad gpa -- criterion variable	gender	prog	rating derived from letters of recommendation	Undergraduate grade point average on 1-9 scale
N	Valid	99	100	98	101	97
	Missing	4	3	5	2	6
Mean		3.3051	1.5200	1.4796	3.6050	6.6959

Correlation

After selecting the variables for the analysis, the specific type of correlation and the type of NHST to be done, the Options window can be used to obtain univariate stats & select the type of Missing Values treatment.



Variables:
- 1st yr grad gpa -- criterion
- gender
- prog
- rating derived from letters
- Undergraduate grade poi

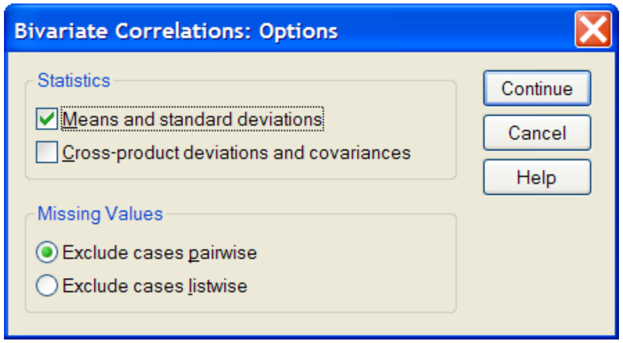
Correlation Coefficients: ☑ Pearson ☐ Kendall's tau-b ☐ Spearman

Test of Significance: ● Two-tailed ○ One-tailed

☐ Flag significant correlations

Pairwise -- each correlation is computed using data from all the participants who have non-missing values for those two variables -- “different samples” representing the population for each correlation but the most “inclusion” for each correlation

Listwise -- all the correlations are computed using only data from participants who have non-missing values for all variables selected -- gives the “same sample” for each correlation, but smallest N



Statistics:
☑ Means and standard deviations
☐ Cross-product deviations and covariances

Missing Values:
● Exclude cases pairwise
○ Exclude cases listwise

Correlation

Pairwise Analysis

Descriptive Statistics

	Mean	Std. Deviation	N
1st yr grad gpa -- criterion variable	3.3051	.61783	99
gender	1.5200	.50212	100
prog	1.4796	.50215	98
rating derived from letters of recommendaton	3.6050	.81183	101
Undergraduate grade point average on 1-9 scale	6.6959	.96436	97

Notice: The gender – ggpa correlation is based on the 96 folks with scores on both, but the gender mean & std are based on N=100 and the ggpa mean & std are based on N=99. Univariate & Bivariate stats are usually not computed from the same participants' data.

Correlations

		1st yr grad gpa -- criterion variable	gender	prog	rating derived from letters of recommendat on	Undergraduat e grade point average on 1-9 scale
1st yr grad gpa -- criterion variable	Pearson Correlation	1	.071	.217	.616	.152
	Sig. (2-tailed)		.491	.036	.000	.072
	N	99	96	94	97	93
gender	Pearson Correlation	.071	1	-.389	-.015	-.071
	Sig. (2-tailed)	.491		.000	.883	.498
	N	96	100	95	98	94
prog	Pearson Correlation	.217	-.389	1	.212	.083
	Sig. (2-tailed)	.036	.000		.038	.219
	N	94	95	98	96	92
rating derived from letters of recommendaton	Pearson Correlation	.616	-.015	.212	1	.198
	Sig. (2-tailed)	.000	.883	.038		.027
	N	97	98	96	101	95
Undergraduate grade point average on 1-9 scale	Pearson Correlation	.152	-.071	.083	.198	1
	Sig. (2-tailed)	.072	.498	.219	.027	
	N	93	94	92	95	97

prog & ggpa → not much difference in r value, but NHST difference (less powerful . . . .
Listwise results are nonsignificant)

ggpa & ugpa → huge difference in r – which one represents the population?

Different correlation results from the two procedures can be because of sample size/power differences, sampling/representation differences, or both.

Listwise Deletion

Descriptive Statistics

	Mean	Std. Deviation	N
1st yr grad gpa -- criterion variable	3.2699	.61302	83
gender	1.5542	.50007	83
prog	1.4819	.50271	83
rating derived from letters of recommendaton	3.5771	.80157	83
Undergraduate grade point average on 1-9 scale	6.6687	.97304	83

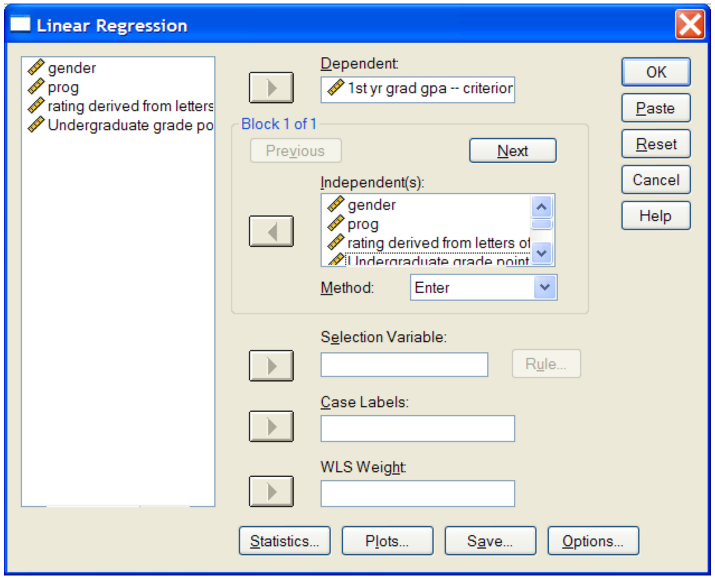
Correlations[a]

		1st yr grad gpa -- criterion variable	gender	prog	rating derived from letters of recommendat on	Undergraduat e grade point average on 1-9 scale
1st yr grad gpa -- criterion variable	Pearson Correlation	1	.035	.202	.614	.642
	Sig. (2-tailed)		.752	.067	.000	.000
gender	Pearson Correlation	.035	1	-.445	.032	-.069
	Sig. (2-tailed)	.752		.000	.774	.534
prog	Pearson Correlation	.202	-.445	1	.224	.330
	Sig. (2-tailed)	.067	.000		.041	.002
rating derived from letters of recommendaton	Pearson Correlation	.614	.032	.224	1	.559
	Sig. (2-tailed)	.000	.774	.041		.000
Undergraduate grade point average on 1-9	Pearson Correlation	.642	-.069	.330	.559	1
	Sig. (2-tailed)	.000	.534	.002	.000	

[a] Listwise N=83

Notice: There were "only a few missing data"(2-6) based on the initial univariate analysis. But if different participants are missing data for different variables, the number lost to Listwise deletion can be substantial.

Multiple Regression



Variables: gender; prog; rating derived from letters; Undergraduate grade po

Dependent: 1st yr grad gpa -- criterior

Independent(s): gender; prog; rating derived from letters of; Undergraduate grade point

Method: Enter

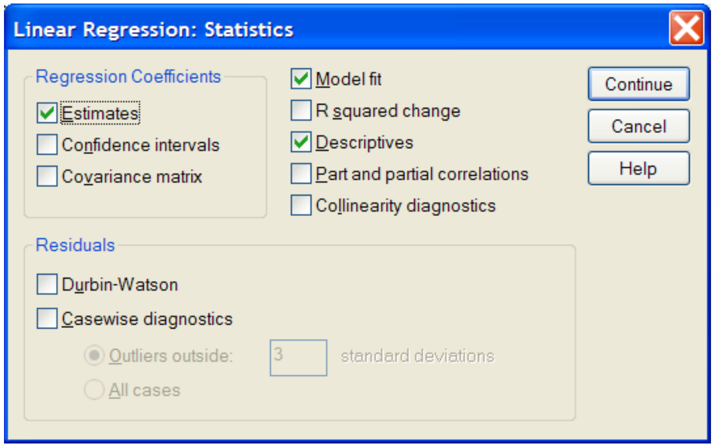
Selection Variable:

Case Labels:

WLS Weight:

Using the Statistics window, you can get univariate statistics and Bivariate correlations. Remember that both of these are calculated as inferential (not descriptive) statistics.

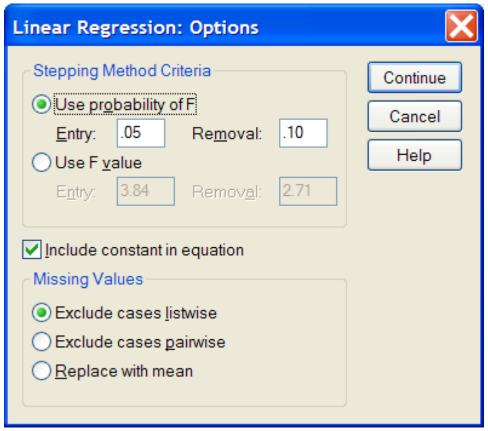
These statistics, as well as the regression model are computed based on the Missing Values procedure chosen from the Options window.



Regression Coefficients
- ☑ Estimates
- □ Confidence intervals
- □ Covariance matrix

- ☑ Model fit
- □ R squared change
- ☑ Descriptives
- □ Part and partial correlations
- □ Collinearity diagnostics

Residuals
- □ Durbin-Watson
- □ Casewise diagnostics
  - ● Outliers outside: 3 standard deviations
  - ○ All cases



Stepping Method Criteria
- ● Use probability of F — Entry: .05 Removal: .10
- ○ Use F value — Entry: 3.84 Removal: 2.71

☑ Include constant in equation

Missing Values
- ● Exclude cases listwise
- ○ Exclude cases pairwise
- ○ Replace with mean

Be sure that the univariate, correlation and multiple regression analyses you report “go together”. It is a good idea to carefully compare the results from separate analyses to be sure you’ve got the right values:

- Compare the mean, stds & Ns obtained via Frequencies, Correlation and Multiple Regression
- Compare the correlations and Ns via Correlation and Multiple Regression

Case wise Deletion

Descriptive Statistics

	Mean	Std. Deviation	N
1st yr grad gpa -- criterion variable	3.2699	.61302	83
gender	1.5542	.50007	83
prog	1.4819	.50271	83
rating derived from letters of recommendaton	3.5771	.80157	83
Undergraduate grade point average on 1-9 scale	6.6687	.97304	83

Note: You'll get the same Casewise correlation matrix as from the Correlation procedure above

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.714[a]	.510	.485	.43976

a. Predictors: (Constant), Undergraduate grade point average on 1-9 scale, gender, prog, rating derived from letters of recommendaton

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	15.730	4	3.933	20.335	.000[a]
	Residual	15.084	78	.193		
	Total	30.815	82			

a. Predictors: (Constant), Undergraduate grade point average on 1-9 scale, gender, prog, rating derived from letters of recommendaton

b. Dependent Variable: 1st yr grad gpa -- criterion variable

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	-.008	.368		-.021	.983
	gender	.162	.097	.132	1.682	.096
	prog	.071	.101	.058	.704	.483
	rating derived from letters of recommendaton	.262	.066	.345	3.967	.000
	Undergraduate grade point average on 1-9 scale	.134	.057	.117	1.234	.101

a. Dependent Variable: 1st yr grad gpa -- criterion variable

The univariate statistics will match those from both the Frequencies and Correlation procedures.

Please Note: The mean, std & N from the Pairwise univariate analyses aren't computed from the same participants as the correlations or the regression model.

As with correlations, different regression results from the two procedures can be because of sample size/power differences, sampling/representation differences, or both.

Note: For the Pairwise Analysis, the df for H0: F-test is based on the smallest pairwise N from the Pairwise correlation.

Pairwise Analysis

Descriptive Statistics

	Mean	Std. Deviation	N
1st yr grad gpa -- criterion variable	3.3051	.61783	99
gender	1.5200	.50212	100
prog	1.4796	.50215	98
rating derived from letters of recommendaton	3.6050	.81183	101
Undergraduate grade point average on 1-9 scale	6.6959	.96436	97

Note: You'll get the same Pairwise correlation matrix as from the Correlation procedure above

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.740[a]	.548	.527	.42477

a. Predictors: (Constant), Undergraduate grade point average on 1-9 scale, gender, prog, rating derived from letters of recommendaton

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	19.038	4	4.760	26.380	.000[a]
	Residual	15.697	87	.180		
	Total	34.736	91			

a. Predictors: (Constant), Undergraduate grade point average on 1-9 scale, gender, prog, rating derived from letters of recommendaton

b. Dependent Variable: 1st yr grad gpa -- criterion variable

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	.314	.400		.785	.435
	gender	.065	.110	.053	.590	.557
	prog	-.004	.115	-.003	-.031	.976
	rating derived from letters of recommendaton	.280	.074	.366	3.800	.000
	Undergraduate grade point average on 1-9 scale	.279	.062	.443	4.482	.000

a. Dependent Variable: 1st yr grad gpa -- criterion variable