This exercise demonstrates relationships between ordinary zero-order correlation, partial correlation, and semipartial correlation. We use hypothetical data representing salaries for 62 faculty members (**salary**), years since PhD (**time**), and number of publications (**pubs**). These data are from Cohen, Cohen, West, and Aiken (2003) Applied Multiple Regression (3rd ed.), Table 3.5.1, p. 83. The data can be downloaded from Sakai (MRC06 ccwaTab3-5-1.sav).

We are interested in the relationship between number of publications and salary, but we are concerned that this relationship is confounded with years since PhD. We will compare simple correlations and partial correlations. Our first step is to examine descriptives and plots.

Click Analyze, Descriptive Statistics, Descriptives…, and select time, pubs, and salary.
Click Options and check Mean, Std. Deviation, Minimum, Maximum, Kurtosis, and Skewness.
Click Continue and Paste to generate the following syntax; run the syntax to produce the table.

```
DESCRIPTIVES
  VARIABLES= time pubs salary
  /STATISTICS=MEAN STDDEV MIN MAX KURTOSIS SKEWNESS .
```

Descriptive Statistics

	N	Minimu	Maximu	Mean	Std.	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Years since PhD	62	1	21	6.79	4.278	1.289	.304	1.661	.599
Number of publications	62	1	69	18.18	14.004	1.377	.304	2.012	.599
Current Salary	62	37939	83503	54816	9706.0	.636	.304	.496	.599
Valid N (listwise)	62								

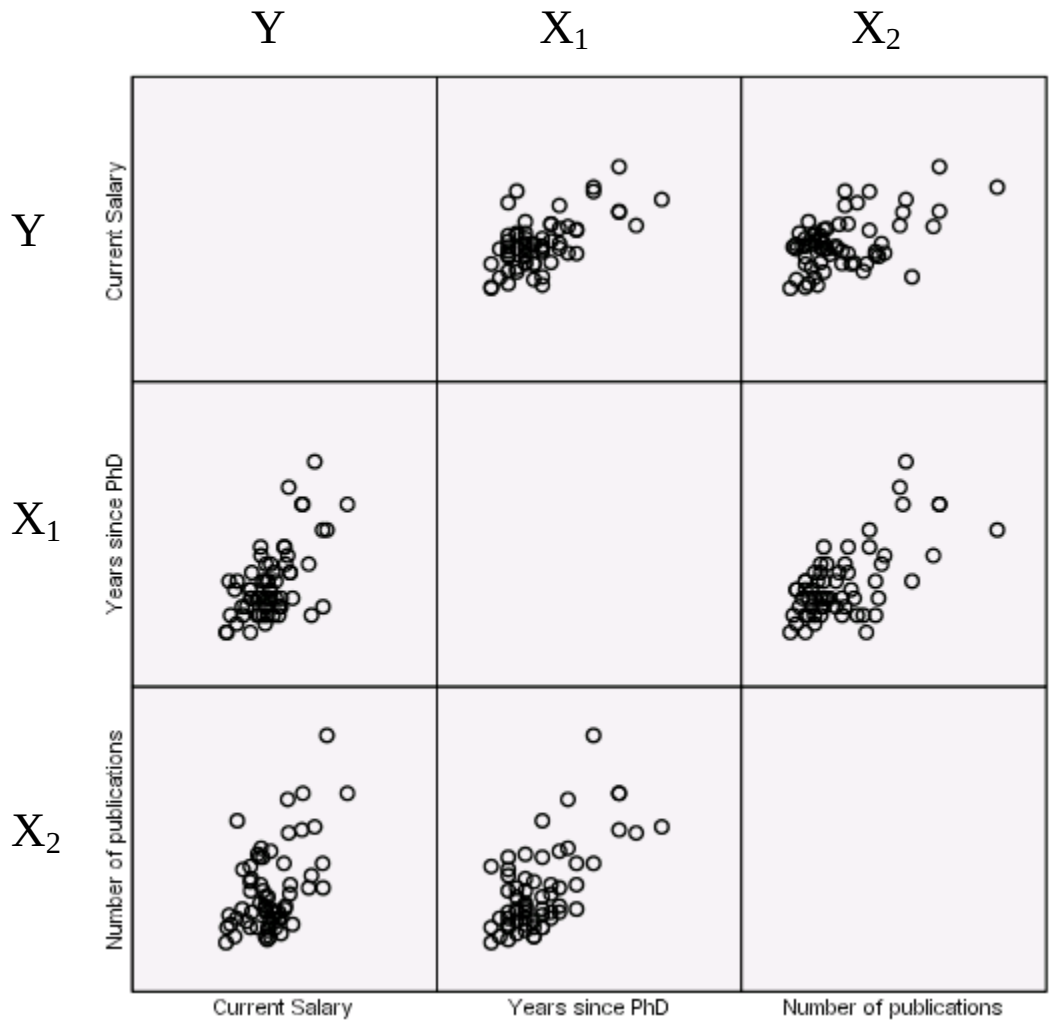
When we examine this table we note that Years since PhD and Number of publications are positively skewed with positive kurtosis greater than 1.0. We should pay special attention to these variables in the plots to make sure we do not have a serious problem with our distributions.

The table of Descriptive Statistics shows only univariate summary statistics. It would be useful to examine univariate plots, too. We also should examine bivariate relationships because we plan to use correlations to measure those relationships. A matrix scatterplot is an efficient way to allow a quick preliminary scan of several bivariate plots simultaneously.

Click Graphs, Legacy Dialogs, Scatter/Dot…, select Matrix Scatter, click Define.
Select the three variables of interest, click Paste to save the syntax shown below:

```
GRAPH
  /SCATTERPLOT(MATRIX)= salary time pubs
  /MISSING=LISTWISE .
```

When we run this syntax we obtain the matrix shown on the next page.



This matrix shows each of the bivariate plots in each orientation. Our primary concern is whether there are any problems in the data that may compromise our analyses. We look for univariate and bivariate outliers, indications of non-linear relationships, and we attend especially to the bivariate relationships with our criterion variable Current Salary where it is shown on the vertical axis (in the first row).

No extreme outliers are apparent, though all variables appear to be positively skewed, especially time and publications. Individual histograms could be useful. A log transform may normalize the distributions, but perhaps the added complication is not justified. We could conduct a 'sensitivity' analyses whereby we redo the analyses with transformed data to see if our conclusions would be materially different.

Let's examine the bivariate correlations. Click Analyze, Correlate, Bivariate…, select the variables **salary**, **time**, and **pubs**. I recommend entering the dependent variable first because that will present all correlations with the dependent variable in the first column.

```
CORRELATIONS
  /VARIABLES=salary time pubs
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE .
```

Correlations

		Current Salary	Years since PhD	Number of publications
Current Salary	Pearson Correlation	1	.608**	.506**
Y	Sig. (2-tailed)		.000	.000
	N	62	62	62
Years since PhD	Pearson Correlation	.608**	1	.651**
X_1	Sig. (2-tailed)	.000		.000
	N	62	62	62
Number of publications	Pearson Correlation	.506**	.651**	1
X_2	Sig. (2-tailed)	.000	.000	
	N	62	62	62

**. Correlation is significant at the 0.01 level (2-tailed).

Note that salary is strongly and statistically significantly correlated with the number of publications (r = .506) and the number of years since PhD (r = .608). Is the relationship between salary and number of publications spurious, in that it could be explained by years since PhD? An example of a spurious relationship is the strong positive correlation that is found between spelling ability of children in grades 1 through 8 and the size of their feet. In that example, age is a confounding variable that has a causal impact on both of the other variables.

We would like to estimate the correlation between salary and the number of publications for people who are equivalent in the number of years since PhD. Would this relationship be much smaller, perhaps zero?

Click Analyze, Correlate, Partial…, select **salary** and **pubs** as the variables and **time** as Controlling for:

```
PARTIAL CORR
  /VARIABLES= salary pubs BY time
  /SIGNIFICANCE=TWOTAIL
  /MISSING=LISTWISE .
```

Partial correlation
$r_{Y2.1} = .184$

Correlations

Control Variables			Current Salary	Number of publications
Years since PhD	Current Salary	Correlation	1.000	.184
		Significance (2-tailed)	.	.157
		df	0	59
	Number of publications	Correlation	.184	1.000
		Significance (2-tailed)	.157	.
		df	59	0

Indeed, we see that when we control for **time** the partial correlation between **salary** and **pubs** is only .184, which does not attain statistical significance. This is our estimate of the correlation between **salary** and **pubs** for a group of people who are all equal on **time**. An assumption is that this correlation is the same at any value of **time**. Thus, if we have a group of people who are all 5

years post-PhD we predict that salary and number of publications would be correlated .184. Ditto for a group of people who are 25 years post-PhD.

Now we will demonstrate that partial correlation is simply the correlation between residuals. We ask SPSS to find and save the residual of **pubs** that cannot be predicted by **time**, and the residual of **salary** that cannot be predicted by **time**. We will see that the correlation between these two residuals is the same as the partial correlation calculated above, .184.

Click Analyze, Regression, Linear…, select **salary** as Dependent and **time** as the Independent. Click Save, check Residuals Unstandardized, and click Continue.

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT salary
  /METHOD=ENTER time
  /SAVE RESID .
```

When we run this syntax, a new variable called **RES_1** is created in our data file. This is the part of **salary** that cannot be predicted with **time**. We could rename this variable **salary.time**.

Now do a comparable analysis where we predict **pubs** using **time**. This creates another new variable in our data file. This one is called **RES_2** and it represents that part of **pubs** that cannot be predicted with **time**. We can rename these variables if we wish, or we could name them as they are created by adding the desired name in parentheses: e.g., **/SAVE RESID (pubs.time) .**

Now let's examine the correlations among all of our variables. Click Analyze, Correlations, Bivariate…, select **salary**, **pubs**, **time**, **RES_1**, and **RES_2**, click Options, select Statistics Means and standard deviations, click Continue, and click Paste. Run the syntax:

```
CORRELATIONS
  /VARIABLES=salary time pubs RES_1 RES_2        *[or salary.time and pubs.time]
  /PRINT=TWOTAIL NOSIG
  /STATISTICS DESCRIPTIVES
  /MISSING=PAIRWISE .
```

Descriptive Statistics

	Mean	Std. Deviation	N	
Current Salary	54815.76	9706.023	62	Y
Years since PhD	6.79	4.278	62	X_1
Number of publications	18.18	14.004	62	X_2
Unstandardized Residual	.0000000	7706.748166	62	← $Y.X_1$ = salary.time
Unstandardized Residual	.0000000	10.63565545	62	← $X_2.X_1$ = pubs.time

Here we see the advantage of providing our own labels in the Variable View window of the SPSS Statistics Editor – the default labels are indistinguishable. Fortunately, we know that the first one is Y.1 or $Y.X_1$ or **salary.time 'salary independent of time'** and the second one is 2.1 or $X_2.X_1$ or **pub.time 'publications independent of time.'**

Y X_1 X_2 $Y.X_1$ $X_2.X_1$

Correlations

		Current Salary	Years since PhD	Number of publications	Unstandardized Residual	Unstandardized Residual
Current Salary	Pearson Correlation	1	.608**	.506**	.794**	.146
Y	Sig. (2-tailed)		.000	.000	.000	.258
	N	62	62	62	62	62
Years since PhD	Pearson Correlation	.608**	1	.651**	.000	.000
	Sig. (2-tailed)	.000		.000	1.000	1.000
X_1	N	62	62	62	62	62
Number of publications	Pearson Correlation	.506**	.651**	1	.139	.759**
X_2	Sig. (2-tailed)	.000	.000		.280	.000
	N	62	62	62	62	62
Unstandardized Residual	Pearson Correlation	.794**	.000	.139	1	.184
$Y.X_1$	Sig. (2-tailed)	.000	1.000	.280		.153
	N	62	62	62	62	62
Unstandardized Residual	Pearson Correlation	.146	.000	.759**	.184	1
$X_2.X_1$	Sig. (2-tailed)	.258	1.000	.000	.153	
	N	62	62	62	62	62

**. Correlation is significant at the 0.01 level (2-tailed).

Semi-partial correlation $r_{Y(2.1)} = .146$.
Note $(.146)^2 = .021$

Partial correlation
$r_{Y2.1} = .184$

Here we see that the correlation between the two residuals is .184, the same value as the partial correlation calculated earlier. (The $p = .153$ is slightly too small because SPSS doesn't know that these variables are residuals – we lost a degree of freedom in the process but didn't tell SPSS.) Note that the first residual (the residual part of **pubs** after **time** has been removed) is not correlated with **time** (r=.000). The percent overlap of the second residual with **pubs** is $(.759)^2 =$ 57.6%, while the percent overlap of **time** with **pubs** is $(.651)^2 = 42.4\%$. Why does this add to 100%? Practice describing and explaining these results.

(The 57.6% represents the portion of pubs that **cannot** be predicted with time, while 42.2% represents the portion of pubs that **can** be predicted with time.)

Next we will create a hierarchical model where we predict **salary** using **time** on the first step and then add **pubs** on the second step. One focus will be the contribution of **pubs** to predicting **salary** beyond the prediction based on **time**. Click Analyze, Regression, Linear…, select **salary** as the Dependent, select **time** as the Independent, click Next, select **pubs** as the Independent, click Statistics, select Model Fit, R squared change, Part and partial correlations, and Estimates. Click Continue and Paste.

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA CHANGE ZPP
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT salary
  /METHOD=ENTER time  /METHOD=ENTER pubs
  /SAVE RESID .
```

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df 1	df 2	Sig. F Change
1	.608[a]	.370	.359	7770.706	.370	35.168	1	60	.000
2	.625[b]	.391	.370	7703.156	.021	2.057	1	59	.157

a. Predictors: (Constant), Years since PhD

b. Predictors: (Constant), Years since PhD, Number of publication

c. Dependent Variable: Current Salary

Semi-partial correlation squared
$(.146)^2 = .021$

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2E+009	1	2123587818	35.168	.000[a]
	Residual	4E+009	60	60383866.76		
	Total	6E+009	61			
2	Regression	2E+009	2	1122820764	18.922	.000[b]
	Residual	4E+009	59	59338615.17		
	Total	6E+009	61			

a. Predictors: (Constant), Years since PhD

b. Predictors: (Constant), Years since PhD, Number of publications

c. Dependent Variable: Current Salary

In SPSS "Part" correlation is semi-partial correlation.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	45450.062	1862.292		24.405	.000			
	Years since PhD	1379.271	232.581	.608	5.930	.000	.608	.608	.608
2	(Constant)	44955.812	1877.993		23.938	.000			
	Years since PhD	1096.027	303.581	.483	3.610	.001	.608	.425	.367
	Number of publications	132.998	92.734	.192	1.434	.157	.506	.184	.146

a. Dependent Variable: Current Salary

Excluded Variables[b]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	Number of publications	.192[a]	1.434	.157	.184	.577

a. Predictors in the Model: (Constant), Years since PhD

b. Dependent Variable: Current Salary

Compare the statistical significance of the R squared added by **pubs** beyond **time** in the Model Summary (this is the square of the semipartial correlation), the unstandardized B weight and the standardized beta coefficients in the final model, and the t-test for the contribution of the excluded variable. Why are these all the same (p = .157)? [They all test the unique contribution of **pubs** in predicting **salary**, controlling for **time**. (or removing **time**, holding **time** constant,. .)]