Chapter 6. Calculating Correlation and Regression in SPSS

Introduction to Correlation and Regression

In lecture, you have learned about two different types of variables; one is the **independent variable (IV)**, and the other is the **dependent variable (DV)**. We study the effects of the independent variable by manipulating it across two or more levels (values). Through our manipulation of the independent variable, we can determine if the independent variable **causes** a change in our dependent variable, our measure. For instance, does varying drug dosage (IV) cause differences in the number of days it takes to recover from a cold (DV).

This type of arrangement allows us to make cause-and-effect statements. If we do see significant differences in our measure, our dependent variable, as a result of testing different values of the independent variable, then we are able to say that it is likely the independent variable caused the change in our measure. You will learn more about these types of manipulations, and the steps to analyze these data in later chapters.

There are times in which we look at variables in a different way; that is, a means other than through an experimental manipulation. We might wish to ask a simpler question to see if and how two variables might be related to one another. In this scenario, we are looking at how two measures change together.

For instance, if we obtained both the heights and weights from a group of people, might we see a relation between these two measures? In general, we know that taller people tend to be heavier than shorter people. While not a perfect relationship, as height increases, weight also tends to increase. This example describes a **positive or direct correlation**. Both measures increase (or decrease) together.

Sometimes we see a relation between measures such that as one measure increases, the other measure tends to decrease. For example, we could ask a group of people how much sleep they had last night, and then obtain a measure of how tired they feel. What might we expect to see? It is likely that a greater amount of sleep would be associated with a lower degree of tiredness. In this case, as hours of sleep increases, the level of tiredness is likely to decrease. The preceding scenario describes a **negative or inverse correlation**; the two measures are associated, and they change together, but in opposite directions.

We can perform **correlational analyses** to determine if there is a significant relationship between our two measures. If we find a significant association does exist, we can then move on to regression analysis. **Regression** procedures allow us to define and plot the best-fitting line through the scatter plot of our correlational data. With linear regression, we are able to generate the formula needed to plot a straight line through the data. Once we have the formula for our regression line, we can use this formula to predict one measure from another; for instance, if we had a measure someone's of height, but not weight, we could use the regression line from our height-weight correlation data to predict that person's weight. We might predict accurately, but we might also be a bit off. How accurately we are able to predict one variable from another depends upon the strength (value) of the correlation.

Data Set 6.1

A principal in a large elementary school is looking over the attendance records of her teachers, many of whom are "working moms." Some have only one child, while others have several children. The principal wonders if having greater numbers of children is associated with increased absence rates. To try answer this questions, she randomly selects the attendance records of 12 of her teachers who are working moms. She then notes how many minor children (i.e., under 18 years old) each teacher has. The data are presented below.

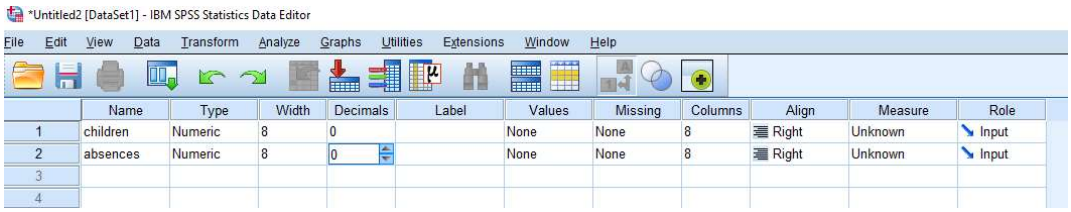
Teacher	children	absences
A	2	5
B	1	3
C	3	9
D	5	13
E	2	8
F	2	4
G	3	3
H	6	10
I	1	6
J	2	6
K	5	7
L	4	6

Correlation Analysis

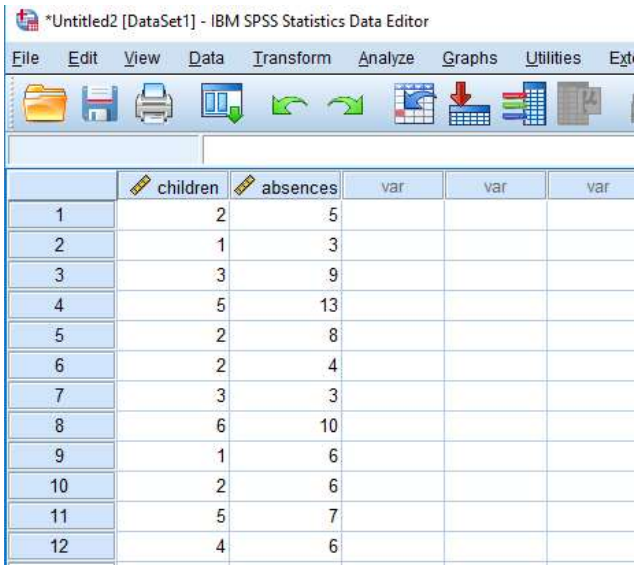
1. Open SPSS, and start with a new data screen.

2. Switch to **"Variable View"** to begin to code your variables.

3. In the **"Name"** column, enter the word **"children"** in the first row, and then enter **"absences"** in the next row. (See below.)

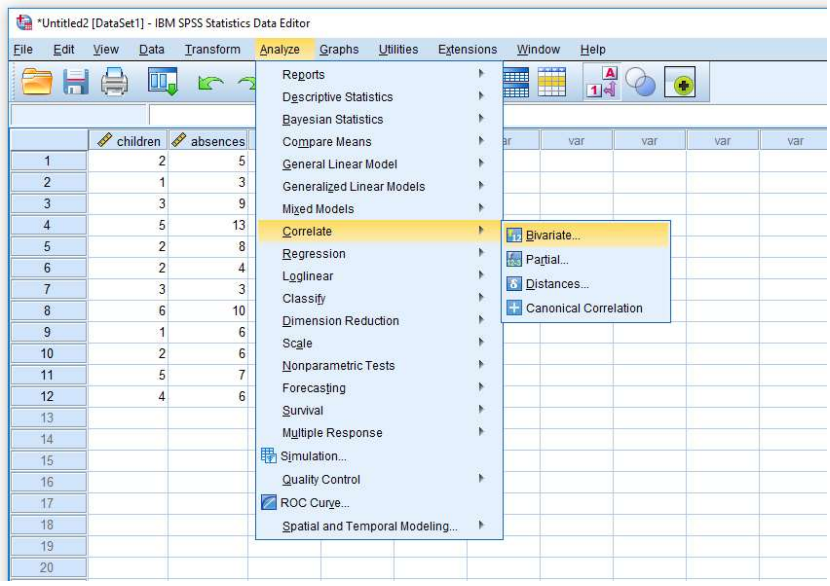


4. Switch to the **"Data View"** tab. Enter the number of children and then the number of absences. Note: each row represents one teacher's data.

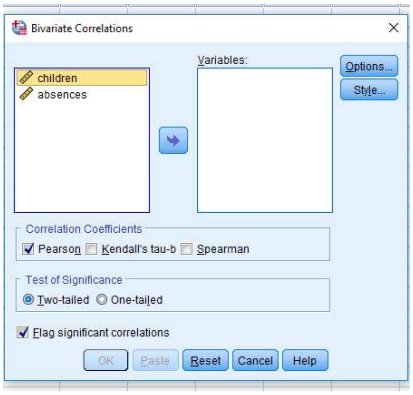


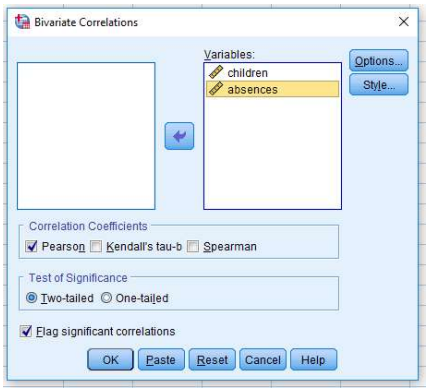
	children	absences	var	var	var
1	2	5			
2	1	3			
3	3	9			
4	5	13			
5	2	8			
6	2	4			
7	3	3			
8	6	10			
9	1	6			
10	2	6			
11	5	7			
12	4	6			

5. Once the data are entered, begin to set up the analysis. To conduct the correlation, go to the **"Data View"** menu bar, and select **"Analyze,"** and then **"Correlate"** followed by **"Bivariate."**



6. The next step is to define the variables. Using the arrow button, first move **"children"** then **"absences"** to the **"Variables"** box. Click **"OK"** to finish the analysis. The results will appear in the SPSS output window.





7. The table below contains our correlation value and the *p* value. The obtained correlation is +.650. (A negative correlation value starts with a minus sign.) The *p* value is .022, and this is **significant** because it is **less than .05.** The degrees of freedom are not shown in the table; however, we can calculate this value using the formula N–2 where "N" is the number of pairs of scores. Here, the degrees of freedom are 10.

Correlations

Correlations

		children	absences
children	Pearson Correlation	1	.650*
	Sig. (2-tailed)		.022
	N	12	12
absences	Pearson Correlation	.650*	1
	Sig. (2-tailed)	.022	
	N	12	12

*. Correlation is significant at the 0.05 level (2-tailed).

Correlation value

p value

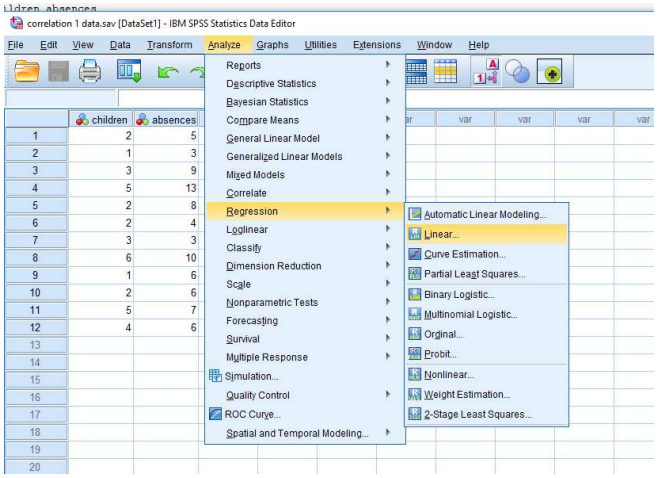
N

8. Now provide an interpretation for the correlation. We use ***"r"*** to represent the correlation. A sample follows.

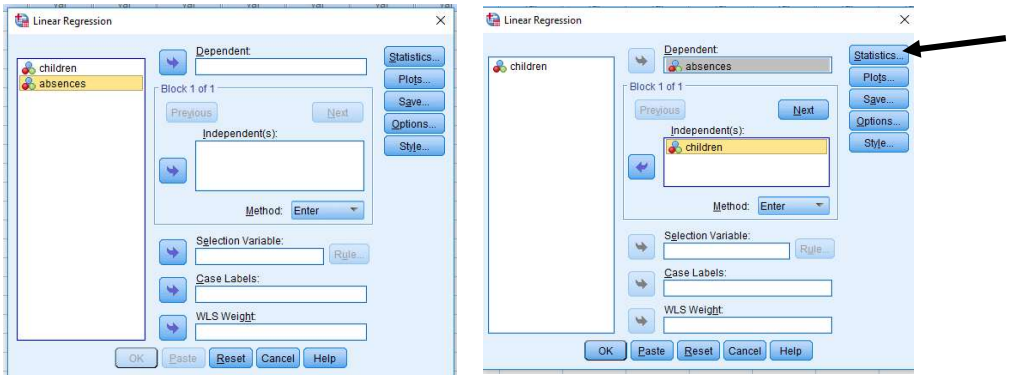
The analysis showed a significant direct (positive) correlation between the teachers' absences and the number of children they have, $r(10) = +.65$, $p = .022$. A greater number of children is associated with a higher absence rate.

9. Since we obtained a significant correlation, we can move on to linear regression analysis.

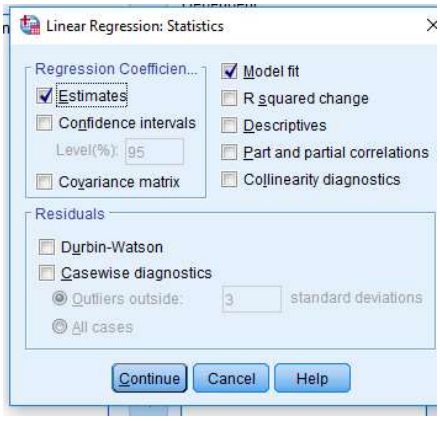
10. First, minimize your SPSS output window, and return to the **"Data View"** tab. From the menu bar, select **"Analyze"** then **"Regression"** followed by **"Linear."** A pop up will appear.



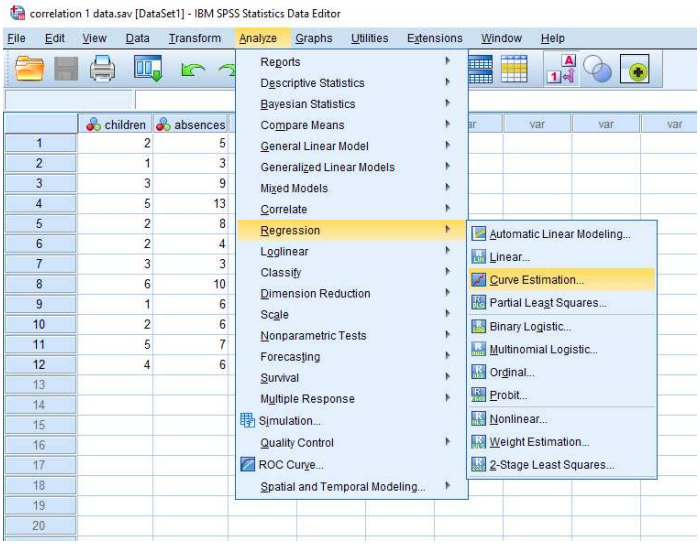
11. The next step is to define the regression variables. Move your **"predictor"** variable, **children**, to the **"Independent"** box, and your **"predicted"** variable, **absences**, to the **"Dependent"** box using the arrow buttons. Once completed, click the **"Statistics"** button.



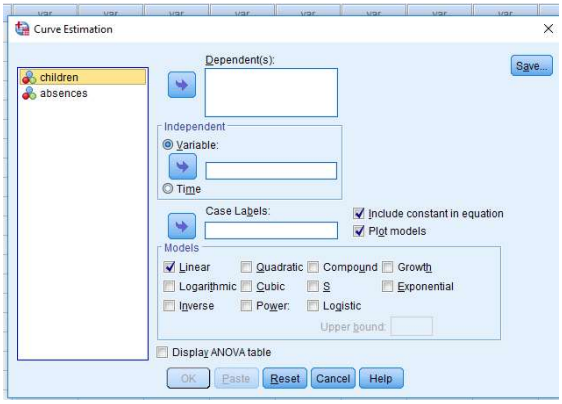
12. A new pop up appears for the Linear Regression. **"Estimates"** and **"Model fit"** must be checked. Click **"Continue"** to proceed.

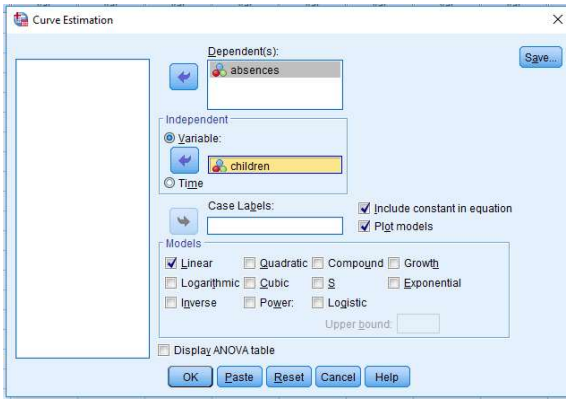


13. From the **"Data View"** menu bar, select **"Analyze," "Regression,"** and then **"Curve Estimation."** These steps are needed to provide the information needed to generate the formula for the regression line.



14. Another pop up appears for the purpose of defining the curve parameters – the slope and y intercept. Move **"absences"** to the **"Dependent"** box, and **"children"** to the **"Variables"** box using the arrow buttons. Click **"OK,"** and the results will appear in the SPSS output window, just below the correlation results.





15. The **"Model Summary"** table gives us our correlation coefficient, or ***"r,"*** as well as ***"r^2,"*** the coefficient of determination. The **"ANOVA"** table shows that we have a **significant** regression, because the *p* value is **less than .05**. The critical table here, is the **"Coefficients"** table. The information for the slope and y intercept for the regression line is presented here.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	children[b]	.	Enter

a. Dependent Variable: absences

b. All requested variables entered.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	40.833	1	40.833	7.313	.022[b]
	Residual	55.833	10	5.583		
	Total	96.667	11			

a. Dependent Variable: absences

b. Predictors: (Constant), children

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.650[a]	.422	.365	2.363

a. Predictors: (Constant), children

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	3.167	1.463		2.165	.056
	children	1.167	.431	.650	2.704	.022

a. Dependent Variable: absences

y intercept

Slope

16. The information provided in the tables above allows for the creation of the regression line.
    - The standard formula for a straight line is
        - **y = bx + a**
    - Since in regression, our "y" values are predicted, and not actual, obtained measures, we note this by putting a **"caret"** mark over the **"ŷ."** It looks like a little hat. Therefore, the formula for a regression line looks a little bit different than that of the standard straight line.
        - **ŷ = bx + a**
            - **x** is the value of the predictor value
                - In this example, it is number of children
            - **ŷ** is the value of the predicted variable
                - In this case, it is number of absences
            - **b** is the slope
                - It measures the rate of change in ŷ as x increases
            - **a** is the y intercept

    - Refer back to the output.
        - The slope is 1.167
        - The y intercept is 3.167
            - Therefore, the formula for the regression line is
                - **ŷ = 1.167x + 3.167**

    - We can now use this formula to predict the number of absences for a mom with a given number of children
        - For example, if a mom has 4 children, we can use the line to predict the number of absence by replacing x with 4.
            - ŷ = 1.167x + 3.167
            - ŷ = 1.167(4) + 3.167
            - ŷ = 4.4668 + 3.167
            - ŷ = 7.6338
            - This tells us that on average, a mom with 4 children is likely to be absent from teaching more than 7 times.

See next page

Practice Problem

Below is a dataset for you to practice correlation and regression analyses.

Your instructor may choose to perform a data collection in class instead, and have you work with those data.

Regardless, perform the correlation analysis, and write an interpretation. If the correlation is significant, complete the regression analysis, and generate the formula for the regression line.

Scenario

A school counselor is worried about her students. She thinks that they might be a bit overcommitted with respect to their responsibilities. Many of them work, and then they have to find time to study around all of their commitments. To better understand if there is a relationship between hours worked, and grades, she gathers a sample of 10 students. She asks how many hours each works per week, and then compares that to their grades on a recent Math exam. The data follow.

Student	Work hours	Grades
A	10	82
B	12	95
C	12	91
D	15	96
E	15	84
F	16	87
G	17	78
H	18	90
I	20	74
J	25	80

- Analyze the data to determine if a significant correlation exists between hours worked and exam grades.
- Write up an interpretation of the findings.
- If there is a significant correlation, conduct regression analyses and determine the formula for the regression line.