

Start of Multi-variate Analysis

➤ *Objectives:*

Showcase the problem of collinearity

Discuss rationale and approach to partial correlations

➤ *Learning Outcomes:*

Become aware of collinearity

Become familiar with partial correlation analysis

Collinearity or Co-linearity

- **Collinearity** is a linear relationship between *two* explanatory variables.
- **Collinearity** of two variables means that strong correlation exists between them, making it difficult (or impossible) to estimate their individual regression coefficients reliably.
- **Multicollinearity** refers to the collinearity of more than two independent variables.
- **Multicollinearity** implies redundancy in the set of variables. This can render ineffective the numerical methods used to solve regression equations.
- Two practical solutions to this problem are to: remove some variables from the model... or combine them.

Multiple Collinearity

Table 1. Correlation coefficients (r-values) for relationships between oceanographic variables sea-surface temperature (SST), sea-surface salinity (SSS), wind speed (WSP), thermocline depth (TDPT), and thermocline slope (TSLP). Sample size (number of 15 min transects) was 2161

	SST	SSS	WSP	TDPT
SSS	-0.588			
WSP	-0.111	-0.056		
TDPT	-0.208	0.083	0.184	
TSLP	0.589	-0.347	0.051	-0.210

(Oedekoven et al. 2001)

Onshore - Offshore:

- Depth
- Distance to shore
- Productivity

North - South:

- Latitude
- Water Temperature
- Wind Speed

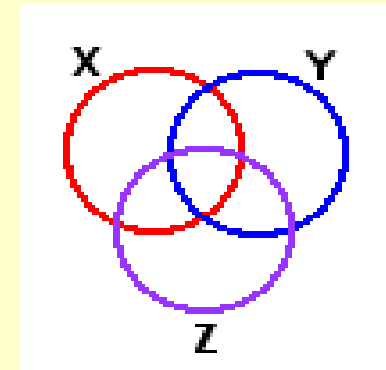
Multiple Collinearity

➤ Suppose you measured N subjects on each of three variables (X, Y, Z) and found the following correlations:

X versus Y : $r_{XY} = +0.50$ $r^2_{XY} = 0.25$

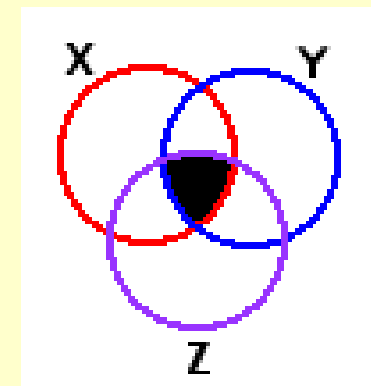
X versus Z : $r_{XZ} = +0.50$ $r^2_{XZ} = 0.25$

Y versus Z : $r_{YZ} = +0.50$ $r^2_{YZ} = 0.25$



➤ The value of r^2 , which equals 0.25, implies that for each pair of variables (XY, XZ, YZ) the covariance, or variance overlap, is 25% (one quarter of the total).

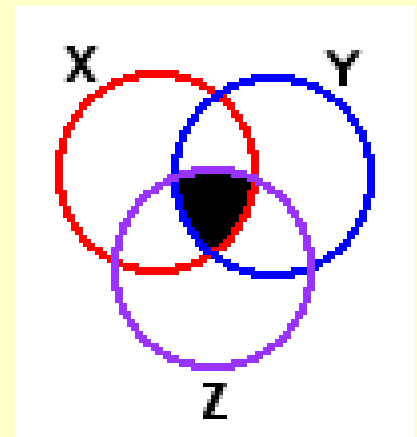
➤ The venn diagram illustrates that 25% of the variability in the three variables (X, Y and Z) overlaps (shaded area)



Partial Correlation - Definition

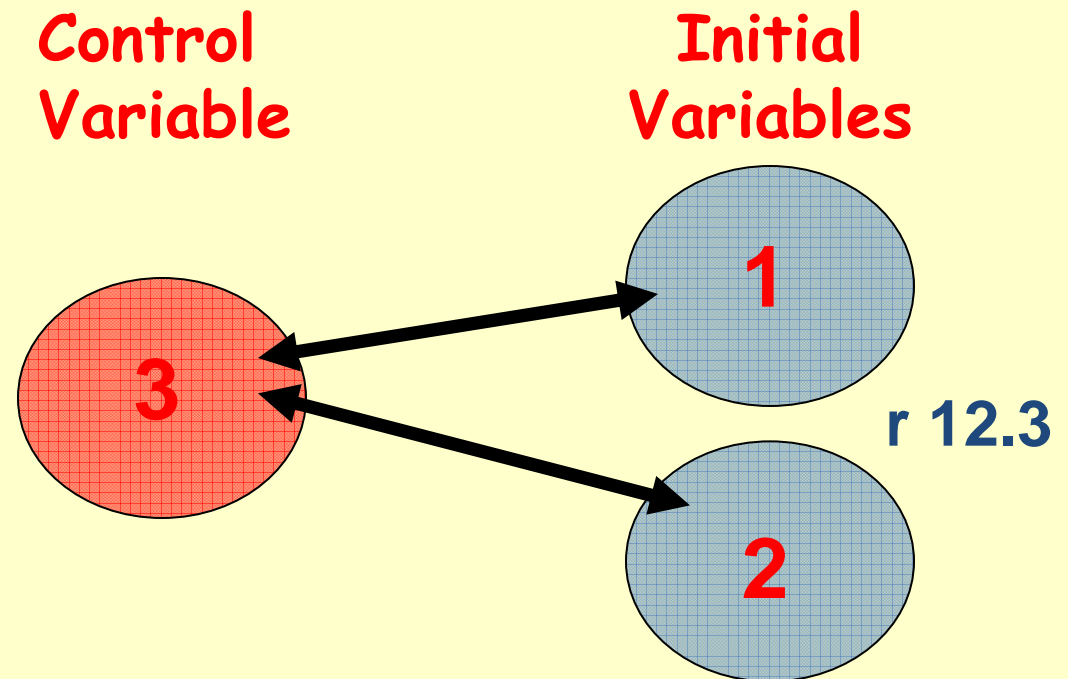
Partial correlation is the correlation of two variables while controlling for a third or more other variables.

- Partial correlation allows us to measure the region of three-way overlap and to remove it from the picture.
- This method determines what value the correlation between any two of the variables would be (hypothetically) if they were not each correlated with the third variable.
- Alternatively, this method allows us to determine what the correlation between any two variables would be (hypothetically) if the third variable were held constant.



Partial Correlation - Application

A control variable is used to extract the variance it explains from each of the two initial variables which are then correlated with each other (variables in zero-order correlation).



The resulting partial correlation ($r_{12.3}$) is the correlation which remains between the two initial variables once the variance explained by the control variable has been removed from each of them.

Partial Correlation - Application

- The technique is commonly used in "causal" modeling of "small models" (with 3 - 5 variables).
- For instance, $r_{12.3}$ is the correlation of variables 1 and 2, controlling for variable 3.
- This approach compares the controlled correlation (e.g., $r_{12.3}$) with the original correlation (e.g., r_{12}) and if there is no difference, the inference is that **the control variable has no effect**.
- If the partial correlation approaches 0, the inference is that the original correlation is spurious - **there is no direct causal link between the two original variables** because the control variables are either (1) common antecedent causes, or (2) intervening variables.

Partial Correlation - Formulation

- The partial correlation of X and Y, with the effects of Z removed (or held constant), is given by the formula:

$$r_{XY \cdot Z} = \frac{r_{XY} - (r_{XZ})(r_{YZ})}{\text{sqrt}[1 - r_{XZ}^2] \times \text{sqrt}[1 - r_{YZ}^2]}$$

which for the present example would work out as

$$\begin{aligned} r_{XY \cdot Z} &= \frac{.50 - (.50)(.50)}{\text{sqrt}[1 - .25] \times \text{sqrt}[1 - .25]} \\ &= \boxed{+.33} \quad (\text{Smaller than } r_{xy} = + 0.50) \end{aligned}$$

Partial Correlation - Formulation

- Partial Correlation of X and Z

$$r_{XZ \cdot Y} = \frac{r_{XZ} - (r_{XY})(r_{YZ})}{\text{sqrt}[1 - r_{XY}^2] \times \text{sqrt}[1 - r_{YZ}^2]}$$

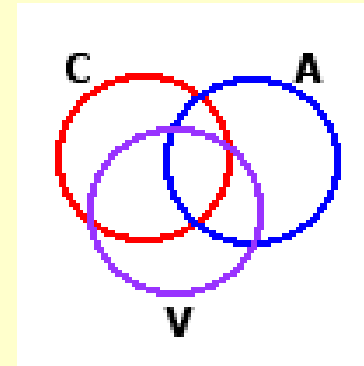
- Partial Correlation of Y and Z

$$r_{YZ \cdot X} = \frac{r_{YZ} - (r_{XY})(r_{XZ})}{\text{sqrt}[1 - r_{XY}^2] \times \text{sqrt}[1 - r_{XZ}^2]}$$

Partial Correlation – Using the Formula

➤ **Example:** Wechsler Adult Intelligence Scale (WAIS) is used to measure "intelligence" beyond the years of childhood. It includes 3 sub-scales labeled C, A, and V.

- C: "comprehension"
- A: "arithmetic"
- V: "vocabulary"

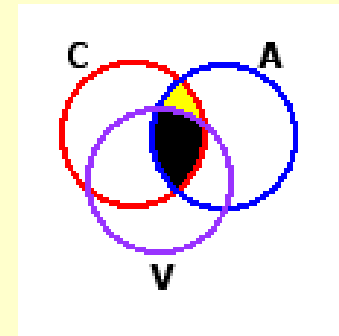
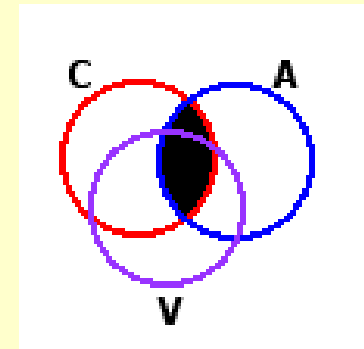


The table shows correlations among these 3 sub-scales:

- C versus A: $r_{CA} = +0.49$ $r^2_{CA} = 0.24$
- C versus V: $r_{CV} = +0.73$ $r^2_{CV} = 0.53$
- A versus V: $r_{AV} = +0.59$ $r^2_{AV} = 0.35$

Partial Correlation – Using the Formula

- While the overlaps are less even, the logic is the same.
- Of the 24% variance overlap in the relationship between comprehension and arithmetic (C & A), a substantial portion reflects the correlations of these variables with vocabulary (V).
- If we remove the effects of V from the relationship between C and A, the partial correlation ($r_{CA.V}$) will be smaller than the full correlation (r_{CA}).



C versus A: $r_{CA} = +0.49$

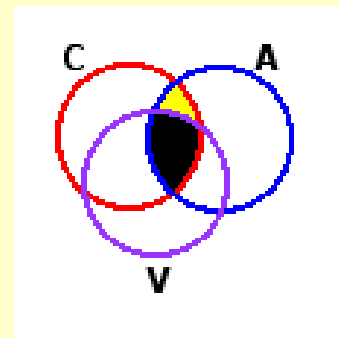
Partial Correlation – Using the Formula

- While the overlaps are less even, the logic is the same.

$$\begin{aligned}r_{CA \cdot V} &= \frac{r_{CA} - (r_{CV})(r_{AV})}{\text{sqrt}[1 - r_{CV}^2] \times \text{sqrt}[1 - r_{AV}^2]} \\ &= \frac{.49 - (.73)(.59)}{\text{sqrt}[1 - .53] \times \text{sqrt}[1 - .35]} \\ &= +.11\end{aligned}$$

$$\text{Hence } r_{CA \cdot V}^2 = .01$$

(Less than $r_{CA} = + 0.49$)



Partial Correlation – Result

Summary: After removing the effects of V , the correlation between C and A diminishes greatly.

- In most cases a partial correlation of the form $r_{XY \cdot Z}$ is smaller than the original correlation r_{XY} .
- In cases where the partial correlation is larger, the third variable, Z , is termed a **suppressor variable** on the assumption that it is suppressing the larger correlation that would appear between X and Y if Z were held constant.

Partial Correlation – Using Regression

- **Example:** Suppose what we want to make good university admissions decisions to maximize our prediction of achievement in college from what we know from the end of high school in the area of mathematics.
- We analyze the data from 1st year college students:
 - SAT (quantitative or math aptitude)
 - CLEP test (math achievement)
 - GPA in first year math sequence

Partial Correlation – Using Regression

Person	SAT-Q	CLEP	Math GPA
1	500	30	2.8
2	550	32	3.0
3	450	28	2.9
4	400	25	2.8
5	600	32	3.3
6	650	38	3.3
7	700	39	3.5
8	550	38	3.7
9	650	35	3.4
10	550	31	2.9

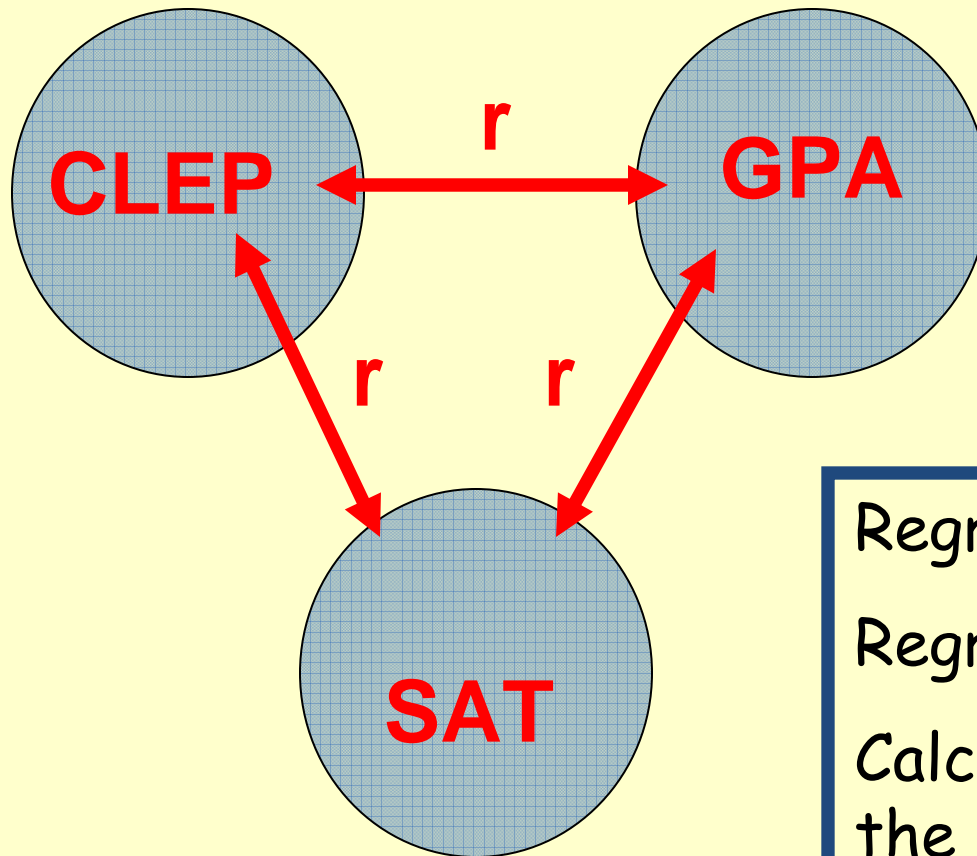
$$df = 10 - 2 = 8$$

$$r \text{ critical} = 0.632$$

	SAT	CLEP	GPA
SAT	-	Sig.	Sig.
CLEP	0.87	-	Sig.
GPA	0.72	0.88	-

Partial Correlation – Using Regression

How to “remove” the effect on one explanatory variable (SAT) to assess correlation between GPA and CLEP?



Regress GPA on SAT

Regress CLEP on SAT

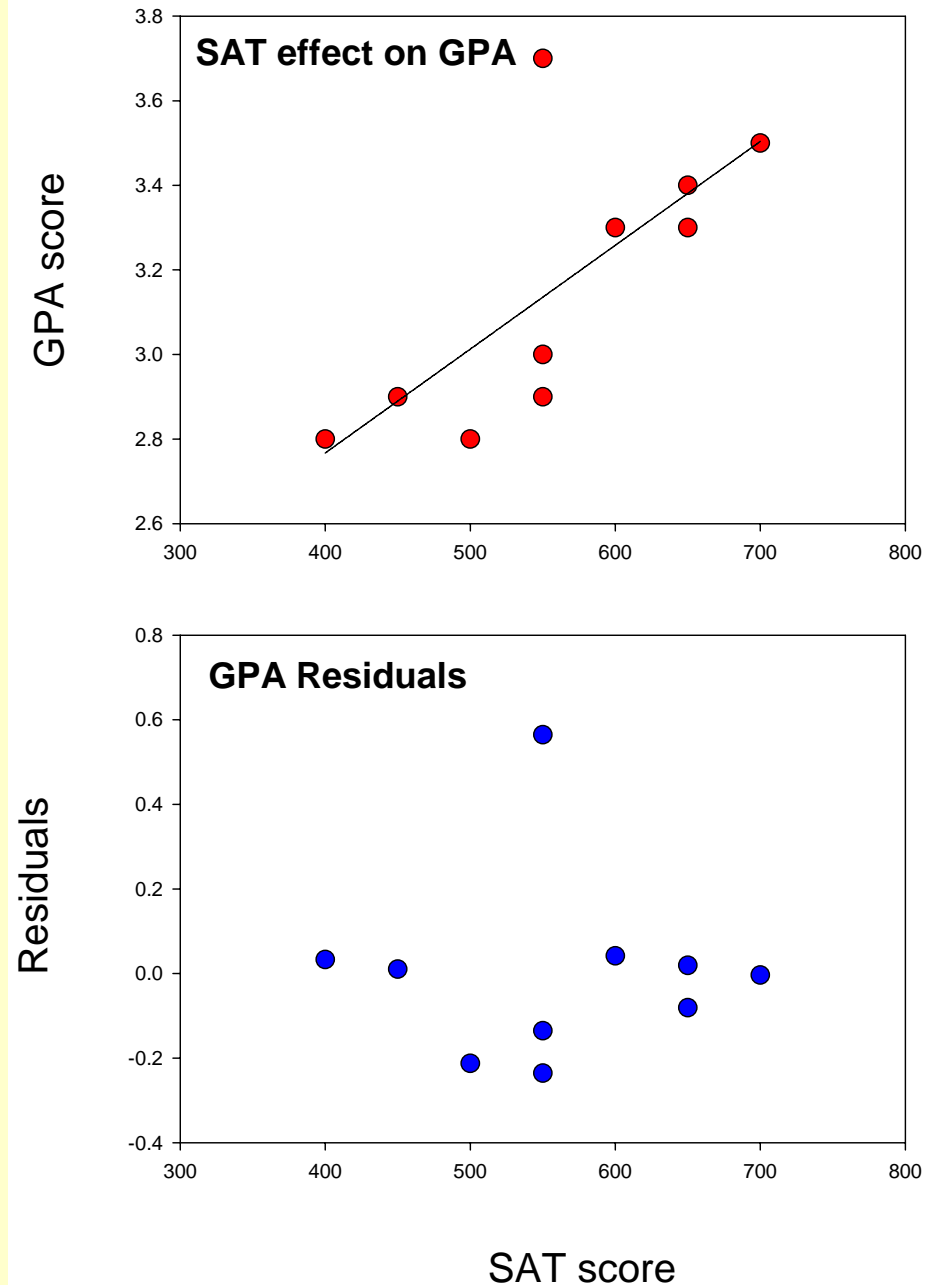
Calculate correlation between the two regression residuals

Linear Regression

$$\text{GPA} = 1.785 + 0.002 \text{ SAT}$$

$$(r^2 = 0.52)$$

Residuals =
Unexplained variation
in GPA, once we have
removed the "linear
influence" of SAT

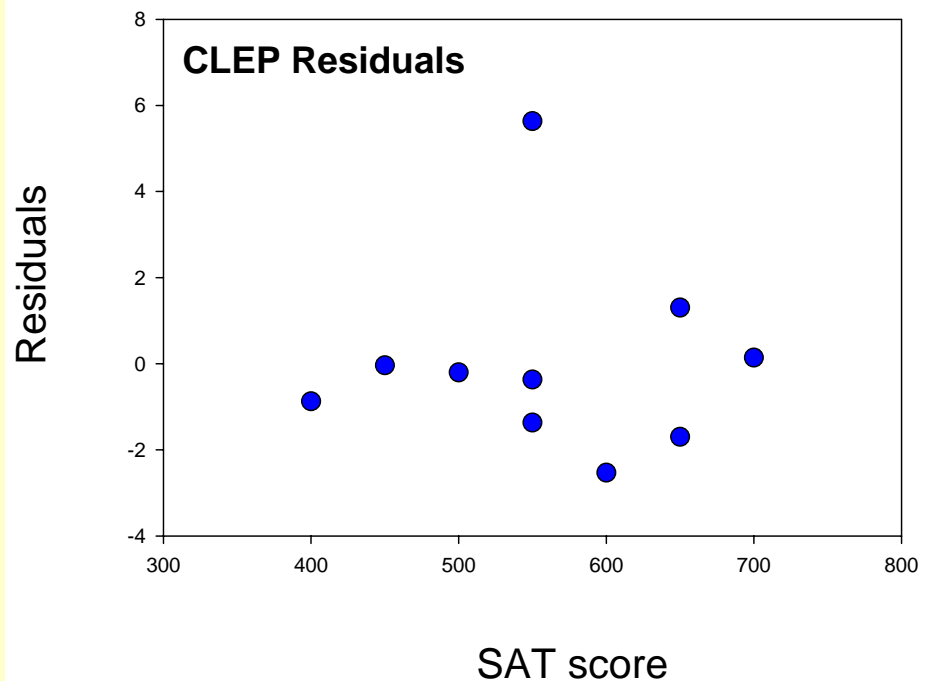
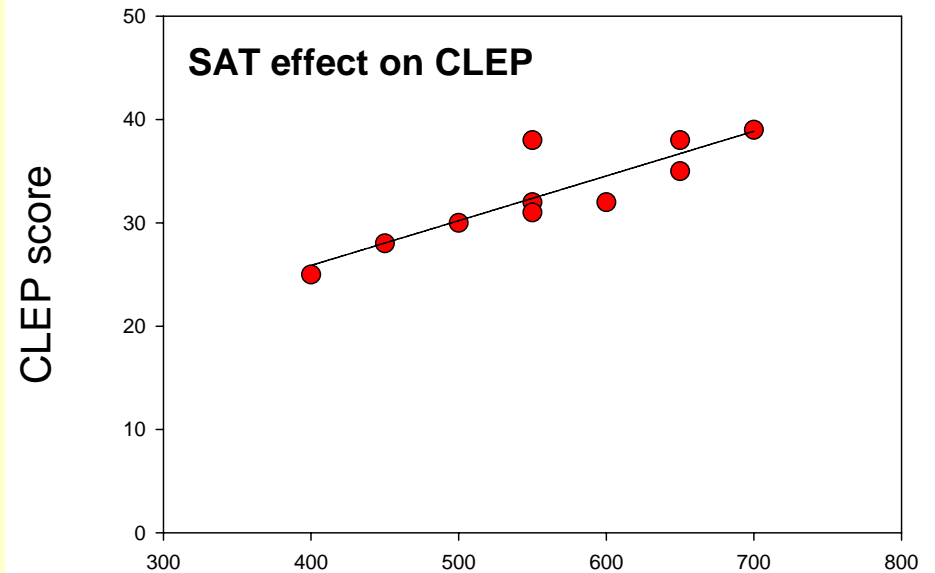


Linear Regression

$$\text{CLEP} = 8.557 + 0.043 \text{ SAT}$$

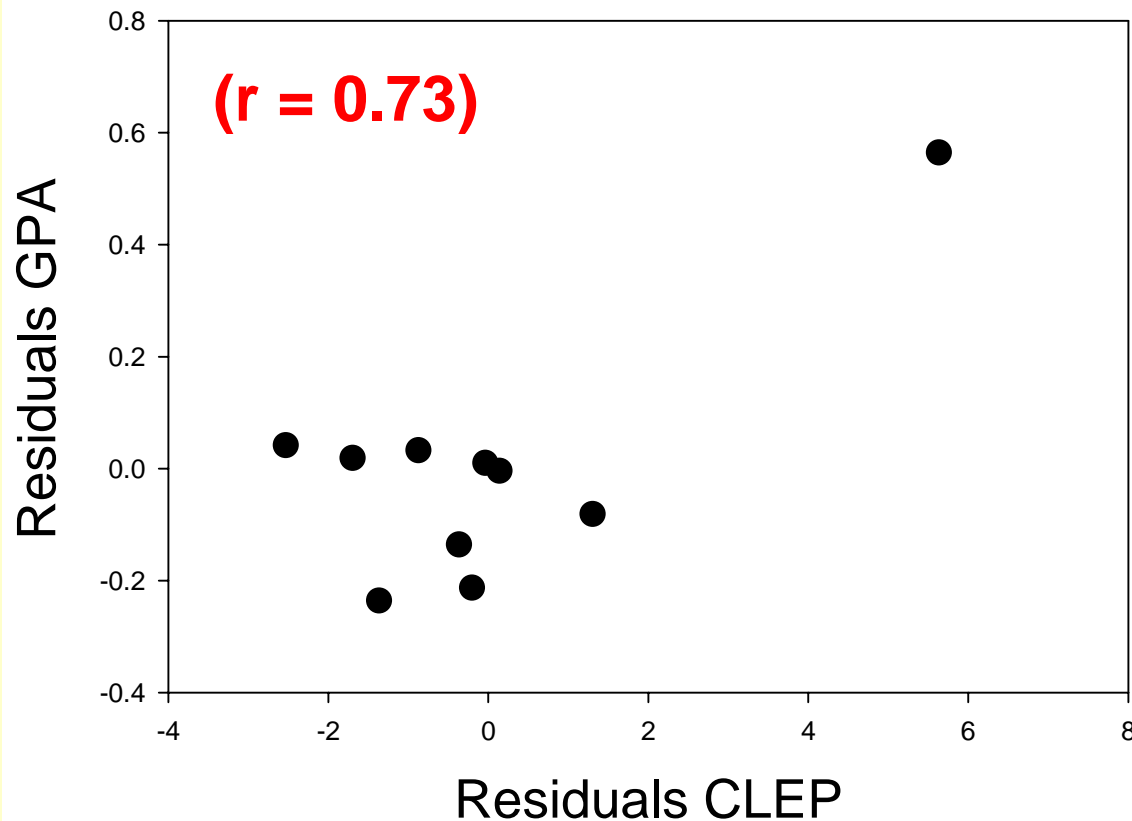
$$(r^2 = 0.76)$$

Residuals =
Unexplained variation
in CLEP, once we have
removed the "linear
influence" of SAT



Partial Correlation – Using Regression

- The correlation between residuals is significant:



The partial correlation is denoted by $r_{12.3}$

r_{12} is the correlation between X_1 and X_2 and .3 means the partial controlling for X_3

Partial Correlation – Comparing Methods

The correlation between GPA and CLEP, while holding SAT constant, is calculated as follows:

The partial correlation is:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2} \sqrt{1 - r_{23}^2}}$$

In this example: (1 = GPA, 2 = CLEP, 3 = SAT)

$$r_{12.3} = \frac{.88 - (.87)(.72)}{\sqrt{1 - .87^2} \sqrt{1 - .72^2}} = .73$$

$$r_{GPA, CLEP.SAT} = .73$$

Partial Correlation – Summary

Order of correlation:

A "first order partial correlation" has a single control variable. A "second order partial correlation" has two control variables ... etc.

A "zero-order correlation" is one with no controls: it is a simple correlation coefficient.

Un-partialled (regular) correlations are *zero order*.

If we partial one variable out of a correlation (e.g., $r_{12.3}$), that partial correlation is a *first order partial correlation*.

If we partial out 2 variables from that correlation (e.g., $r_{12.34}$), we have a *second order partial correlation*

... and so forth.

Partial Correlation – Summary

➤ We can use a formula to compute first order partials, or we can use simple regression to compute the residuals - which are correlated.

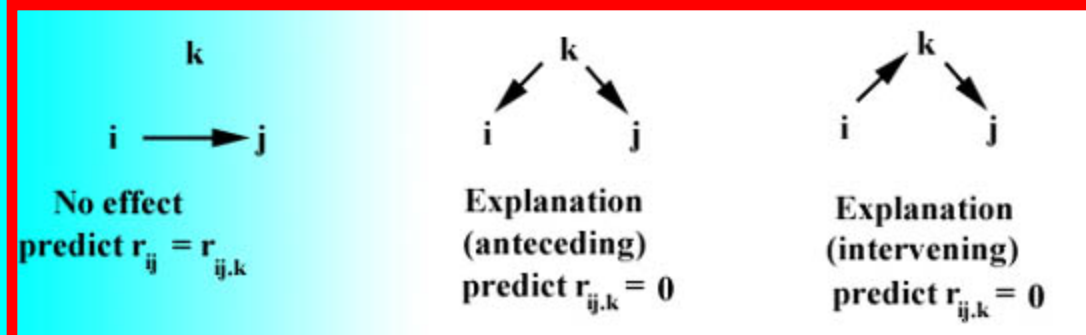
$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2} \sqrt{1 - r_{23}^2}}$$

- For example, to compute $r_{12.3}$, we regress X_1 and X_2 on X_3 and then compute the correlation between the residuals.
- This approach would compute the correlation between X_1 and X_2 , controlling for the influence of X_3 on both.

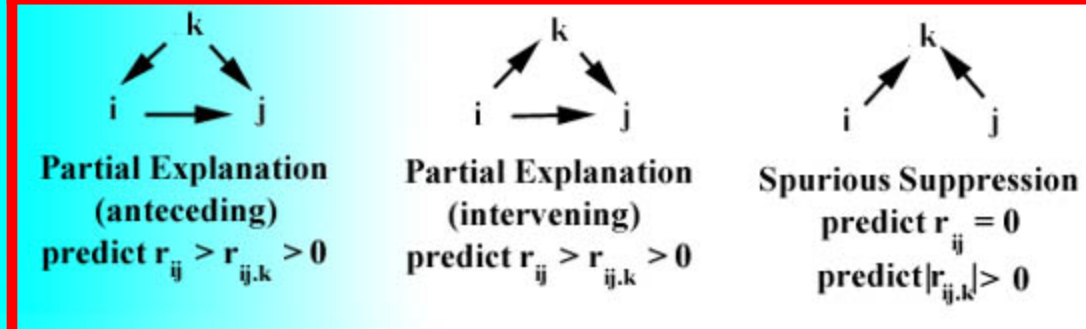
Partial Correlation – Summary

Causal Inference with Partial Correlation

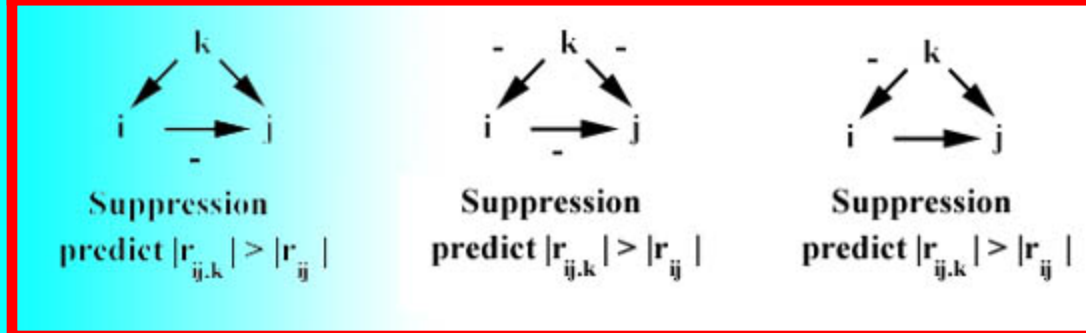
No
Partial
Effect



Partial
Effect



Suppression
Effect



Partial Correlation – Summary

- We can also use a formula to compute second and higher order partials, or we can use multiple regression to compute the residuals - which are then correlated.
- For example, to compute $r_{12.34}$, we could regress each of X_1 and X_2 on both X_3 and X_4 simultaneously and then compute the correlation between the residuals.
- This approach would compute the correlation between X_1 and X_2 , controlling for the influence of both X_3 and X_4 .