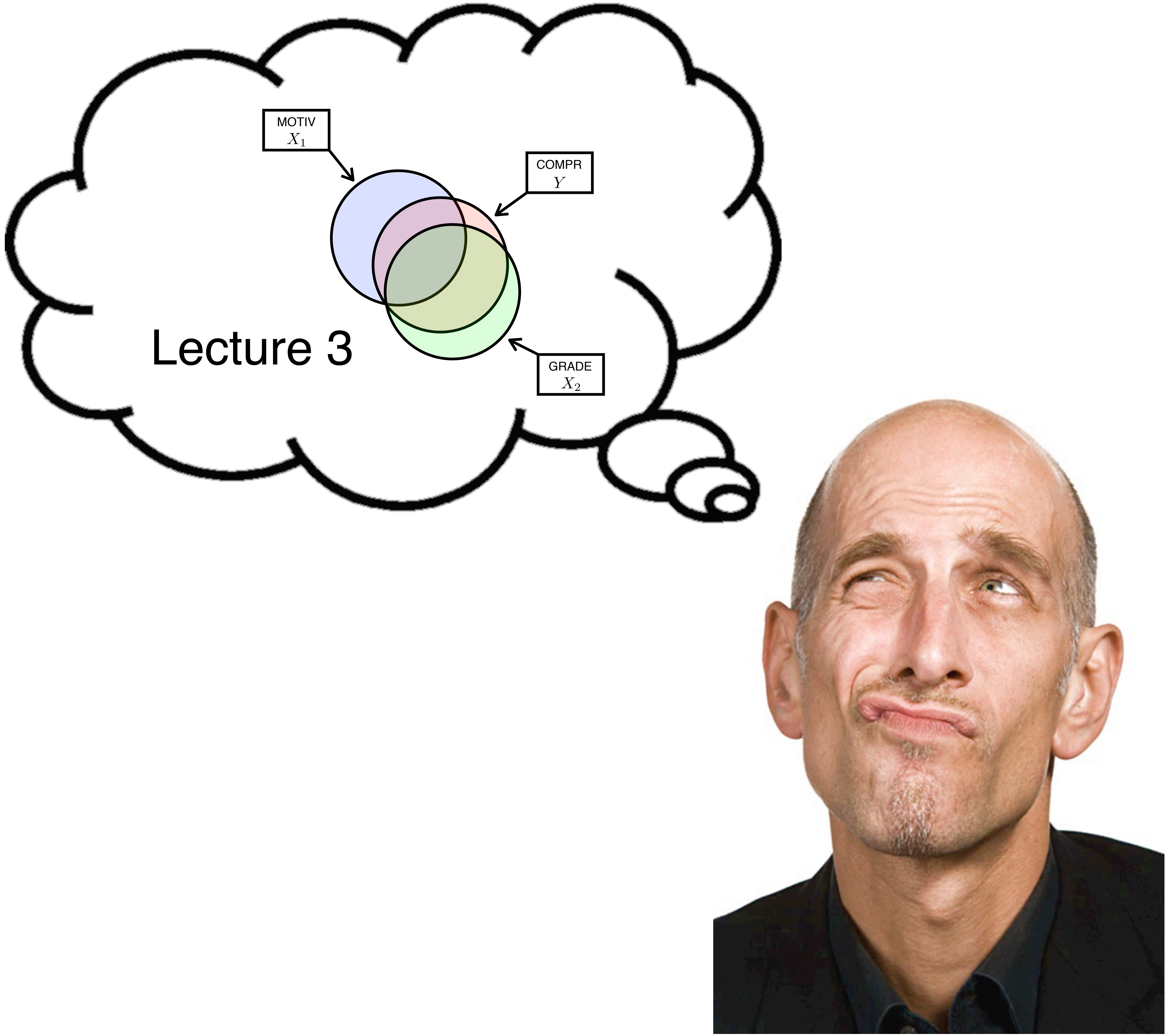
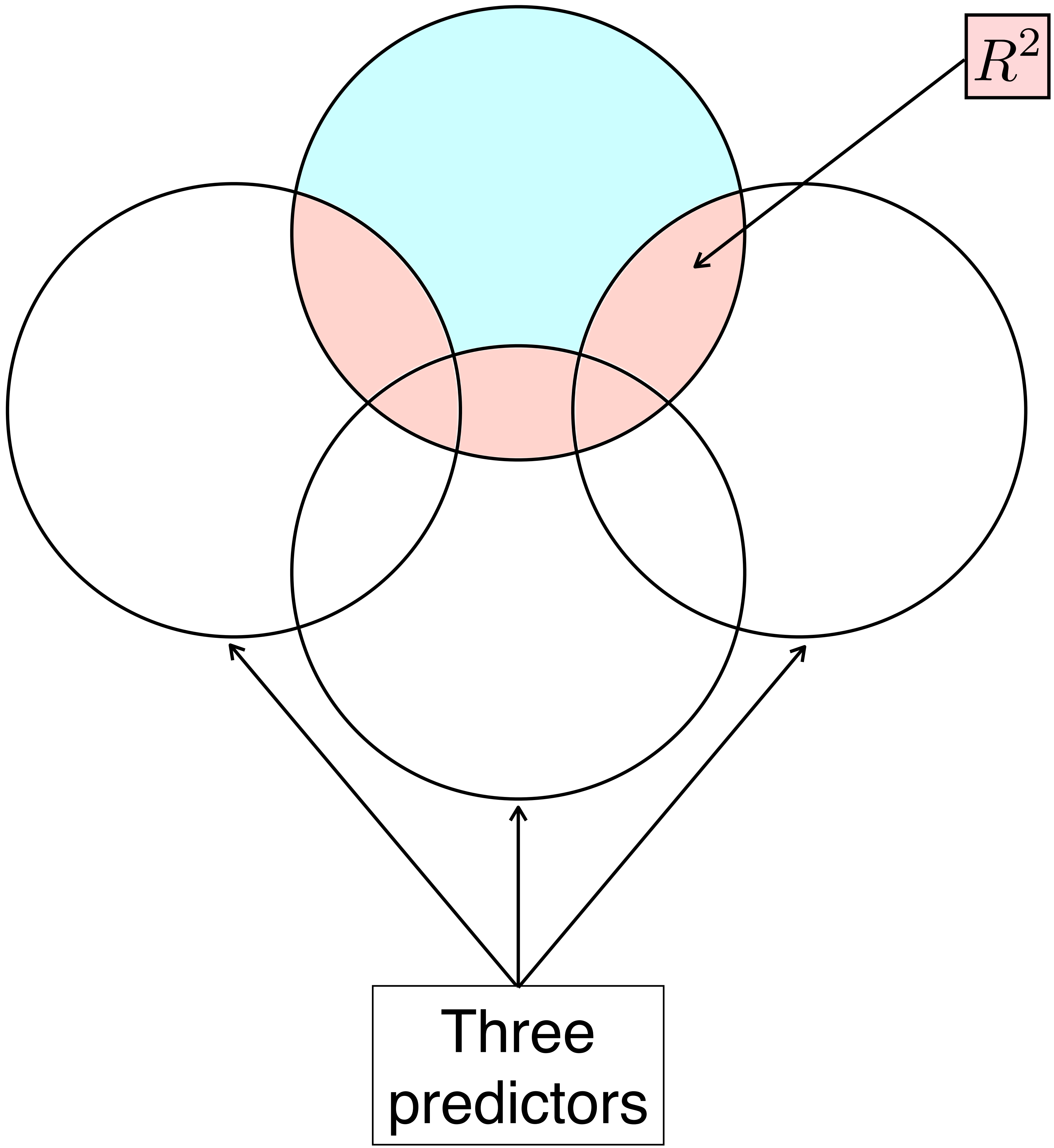
Questions (Data Checking)

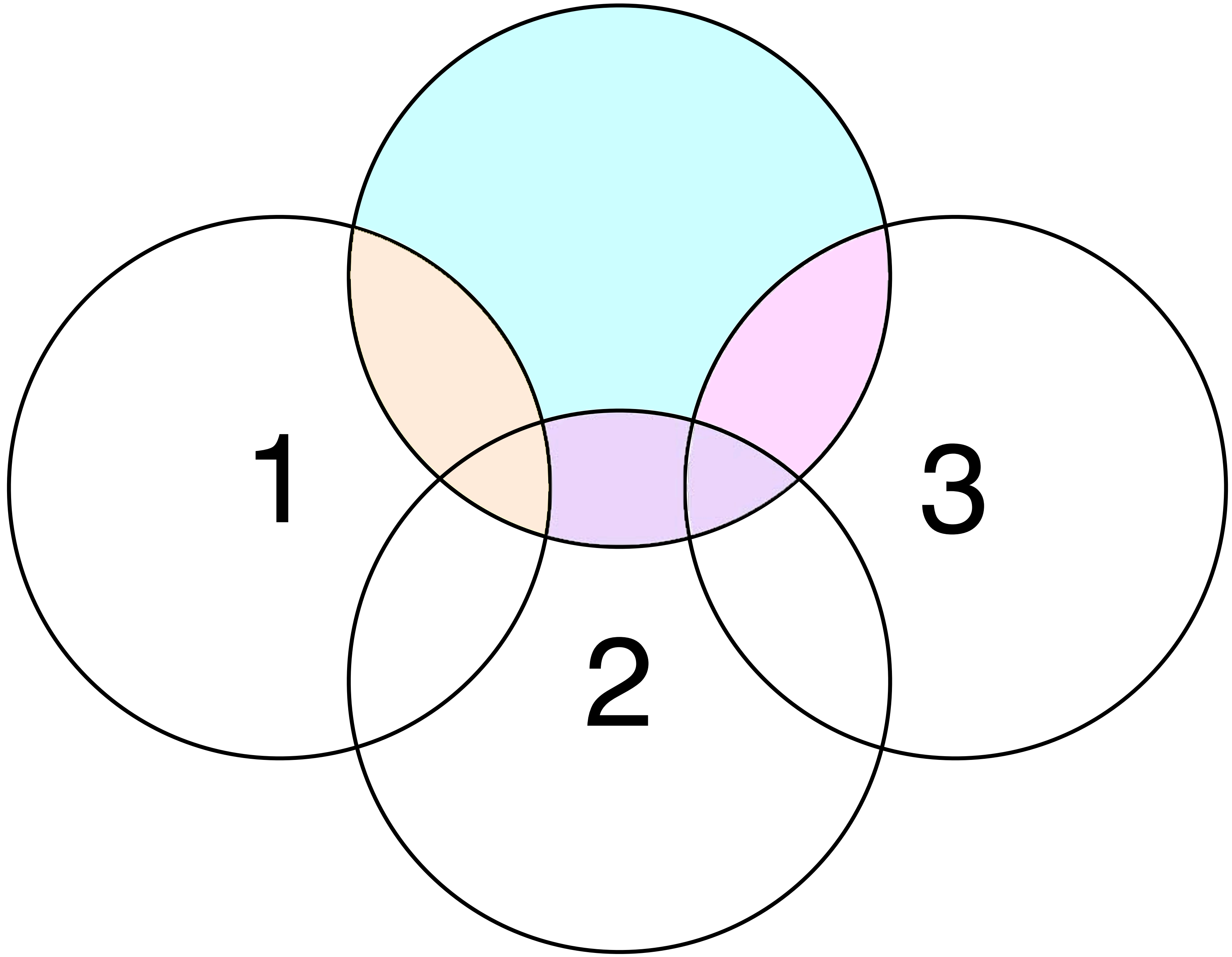
1. Why is data checking important?
2. What are the strengths and weaknesses of using graphical methods to examine data?
3. What are the strengths and weaknesses of using statistical methods to examine data?
4. Why should care be taken when interpreting correlation coefficients?
5. List potential underlying causes of outliers.
6. Discuss why outliers might be classified as beneficial and as problematic.
7. Discuss the following statement: multivariate analyses can be run on any data set, as long as the sample size is adequate.

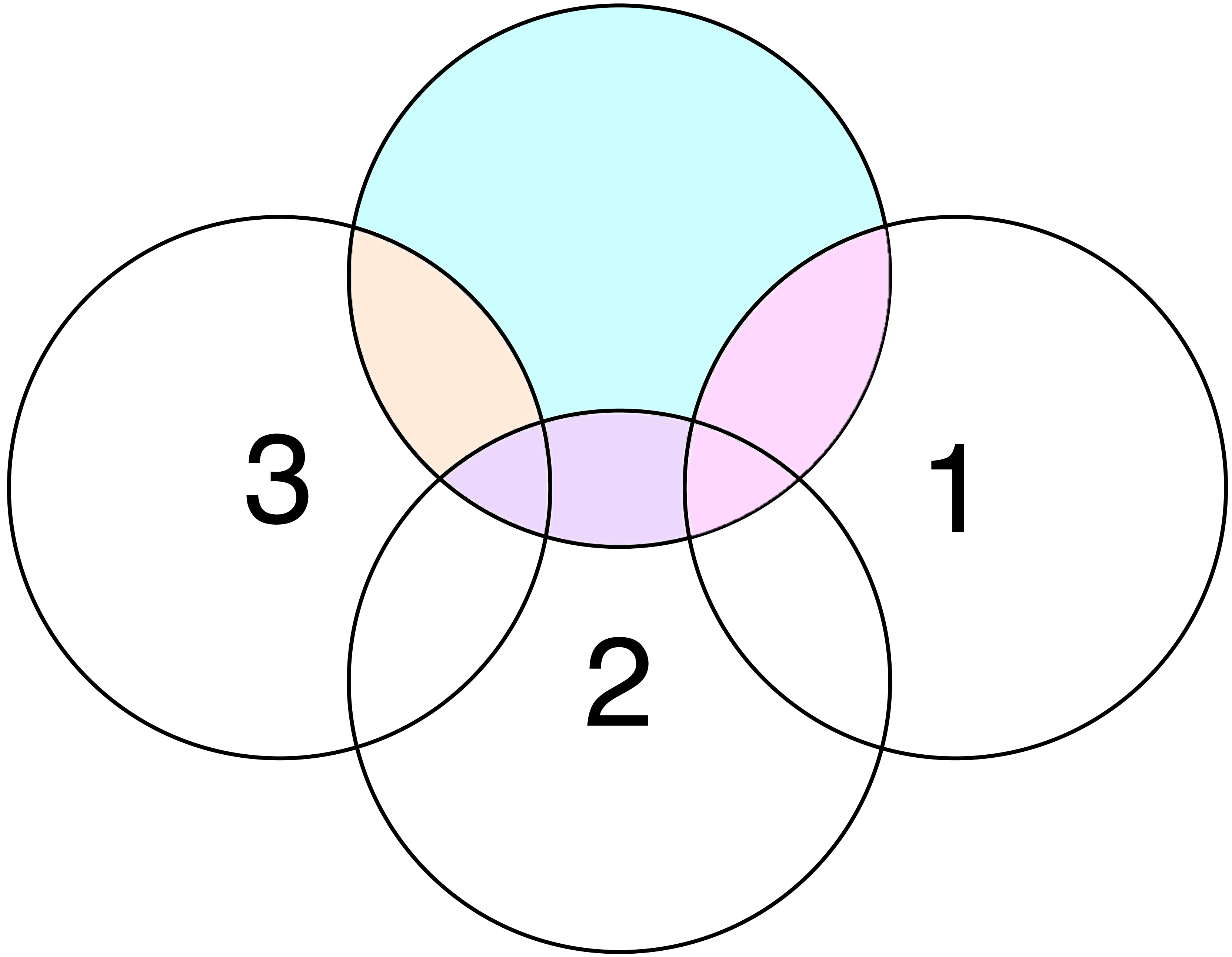
Questions (Diagnostics)

1. Why is it important to examine the assumption of linearity when using regression? How can nonlinearity be corrected or accounted for in the regression equation?
2. Are influential cases always to be omitted? Give examples of occasions when they should or should not be omitted.
3. Describe the reasons for not relying solely on the univariate correlation matrix for diagnosing multicollinearity.
4. How is data checking related to statistical assumptions? How is data checking evaluated?
5. Explain the role judgement calls and ethics play in data diagnostics.


MOTIV
X_1
COMPR
Y
GRADE
X_2
Lecture 3


R^2
Three
predictors


1
2
3

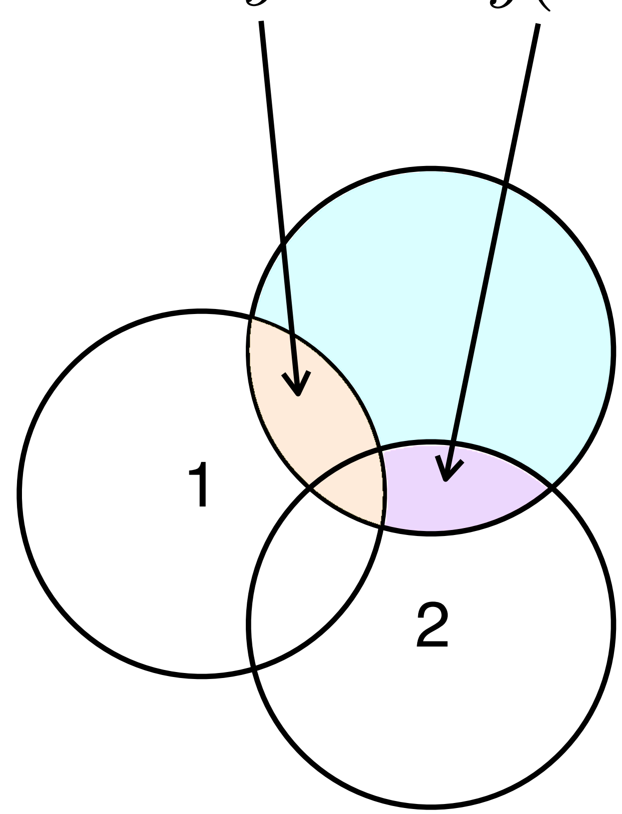


Altering the order in which the variables is the basis for Sequential Multiple Regression

Independent effects of each predictor

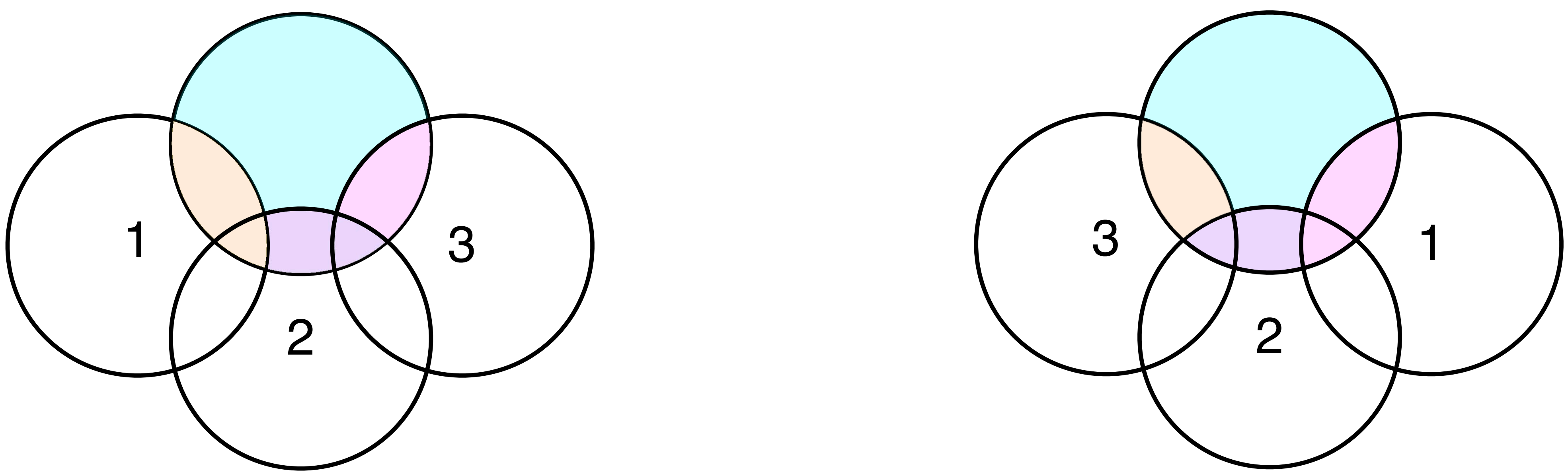
- R^2 can be expressed in terms of independent effects:

$$R^2 = r^2_{y1} + r^2_{y(2 \cdot 1)}$$



Independent effects of each predictor

- R^2 can be expressed in terms of independent effects:

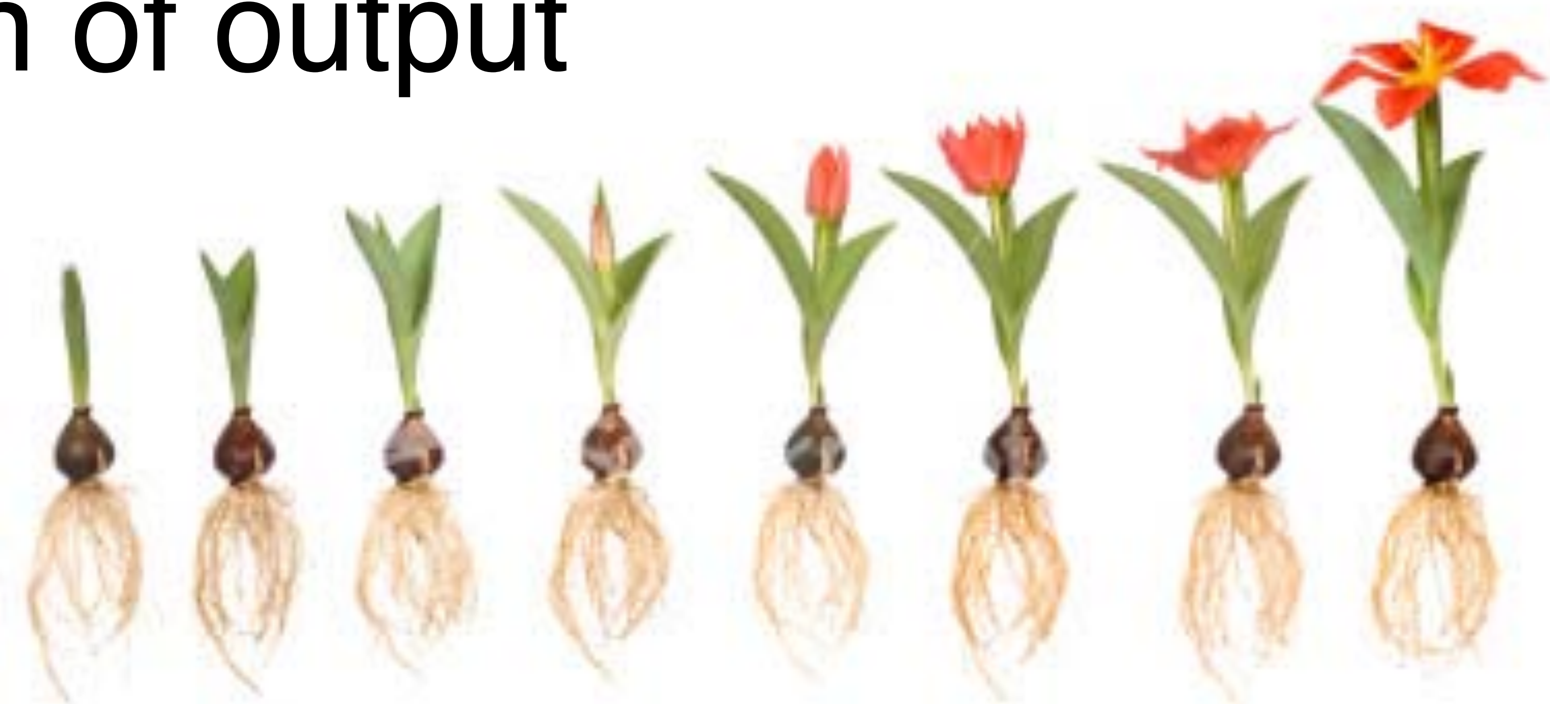


$$R^2 = r^2_{y1} + r^2_{y(2\cdot 1)} + r^2_{y(3\cdot 21)}$$

Expressing R^2 in this way indicates that changing the order of the predictors changes their relative importance.

Sequential (hierarchical) Regression

- Overview
- An example
- SPSS commands
- Interpretation of output



Sequential Regression: An Overview

Rather than being entered all at once, predictors enter the equation in groups specified by the researcher.

The order of entry comes from:

- logical or theoretical considerations
- wanting to co-vary out the effects of certain variables

Each group or block of predictors is assessed in terms of what additional variance it explains.

As an example consider the regression model...

$$Y \leftarrow X_1, X_2, X_3, X_4, X_5, X_6, X_7$$

X_3, X_4, X_5 are thought to be ‘covariates’ or perhaps ‘nuisance variables’.

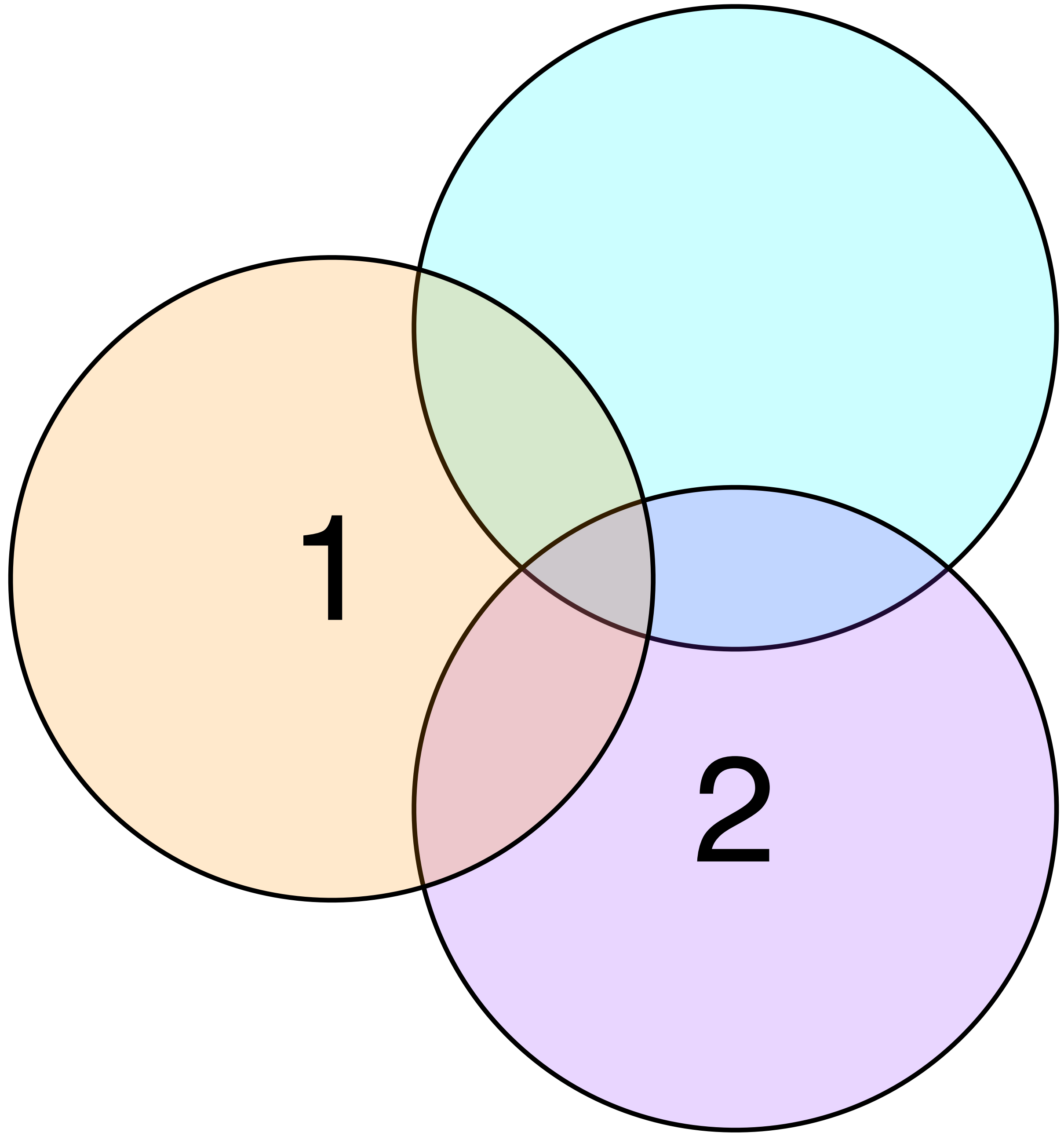
We may be interested in the importance of variables X_1, X_2, X_6, X_7 after X_3, X_4, X_5 have been taken into account.

Formulae for the change in R^2

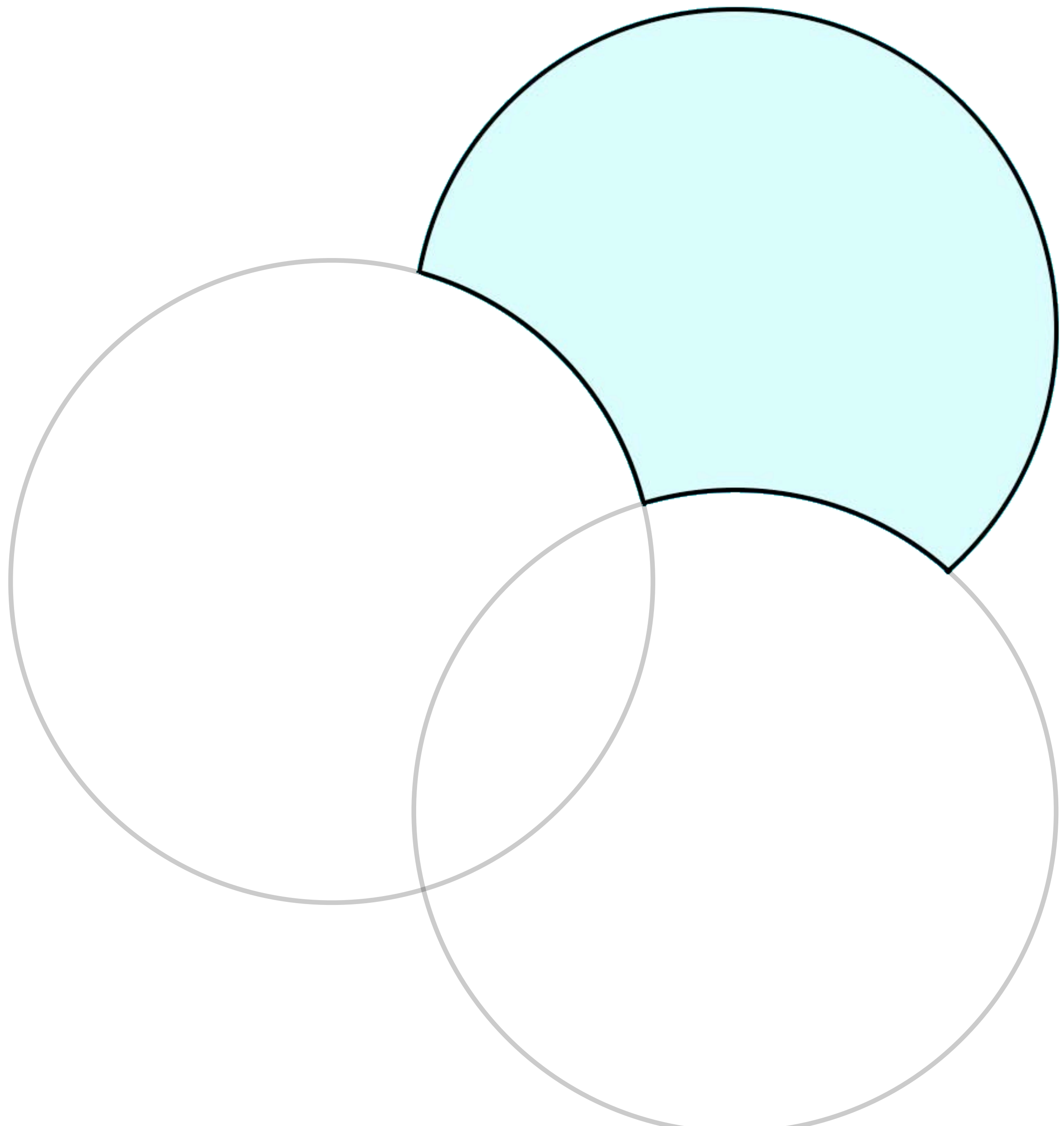
$$R^2_{change} = R^2_{y \cdot 1234567} - R^2_{y \cdot 345}$$

$$R^2_{change} = R^2_{full} \qquad -R^2_{reduced}$$

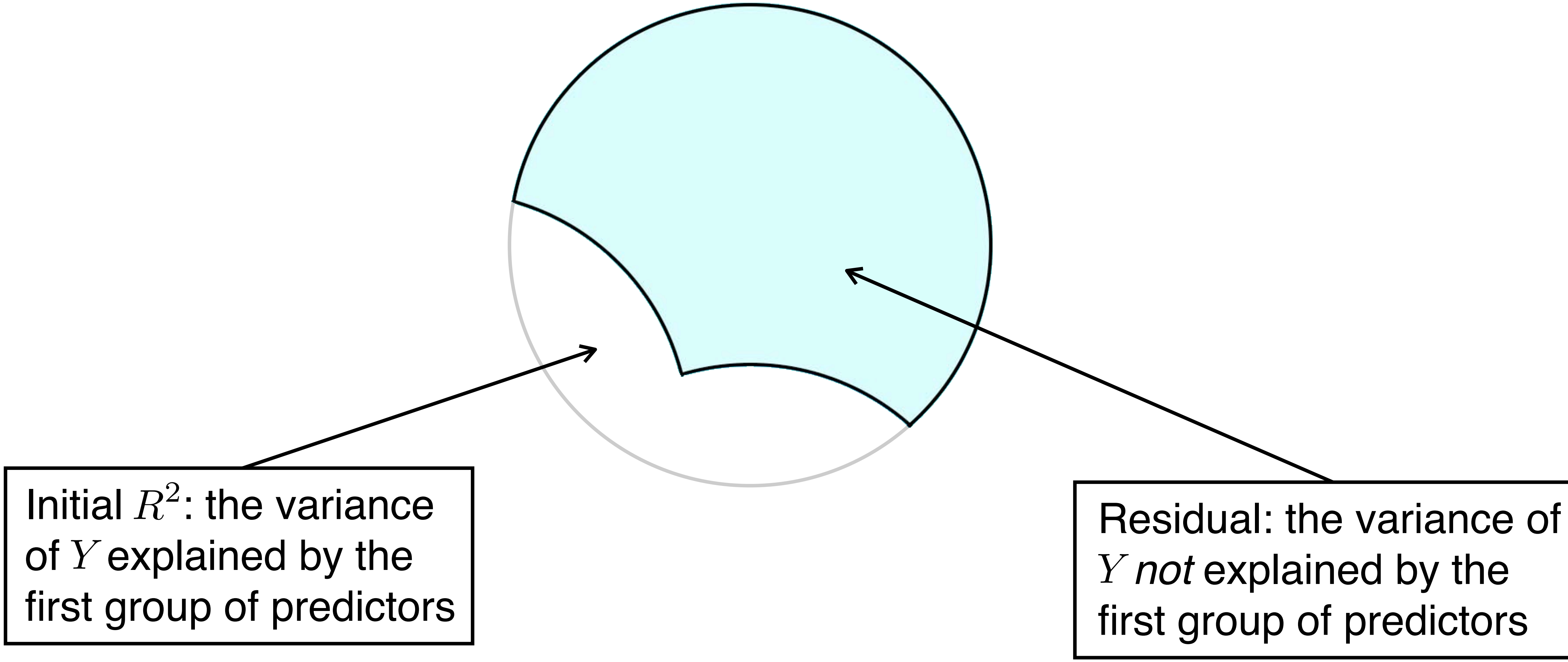
R^2_{change} can be tested for significance using an F-statistic (as per normal R^2)



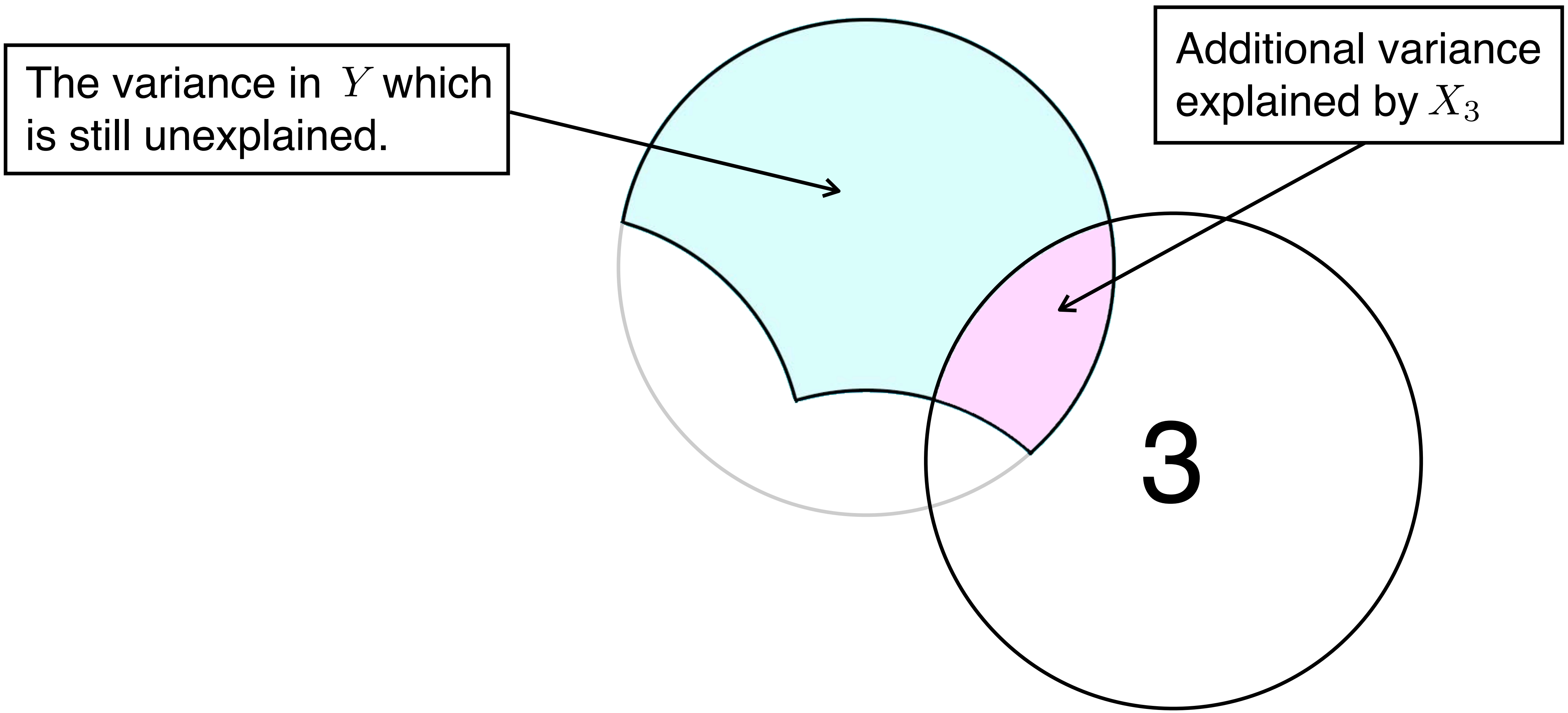
Consider X_1, X_2 and Y



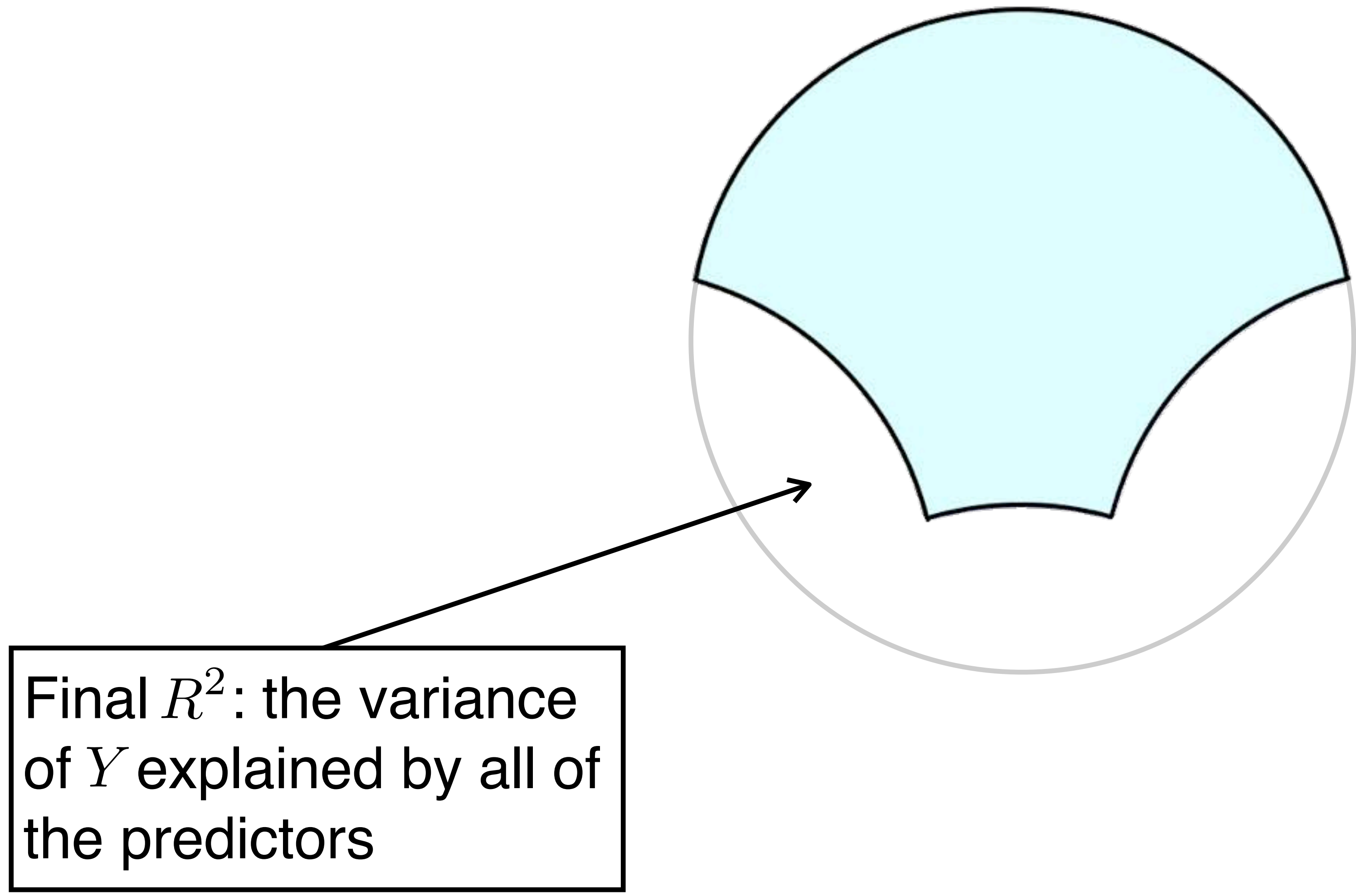
Y with X_1 and X_2 removed



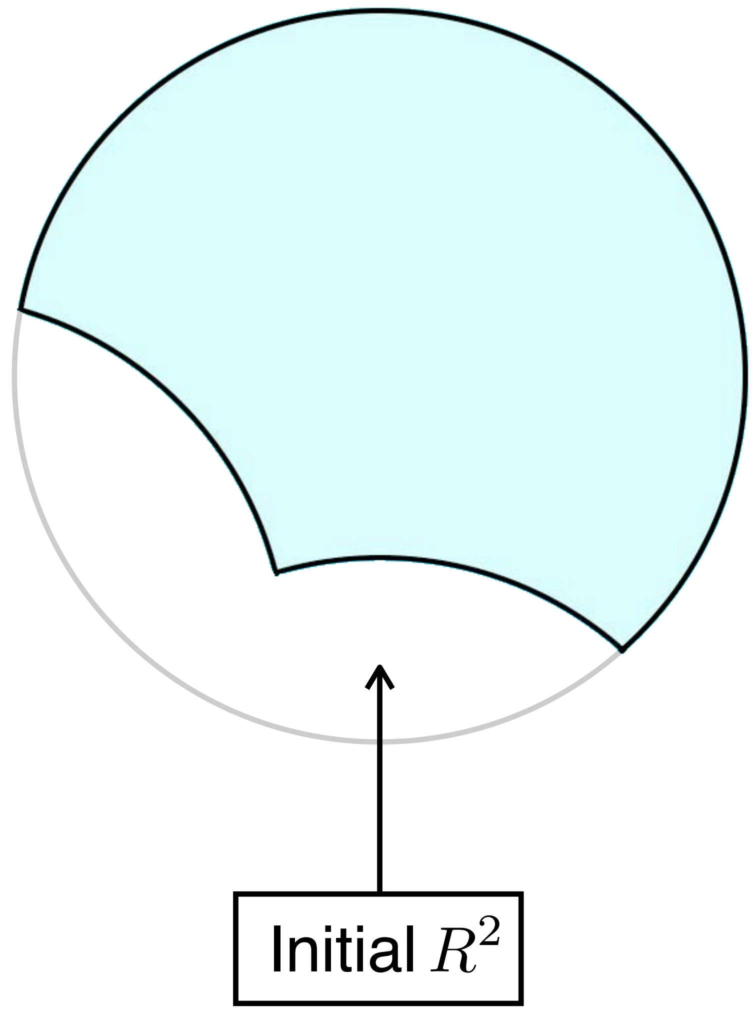
Explained and unexplained sections of Y

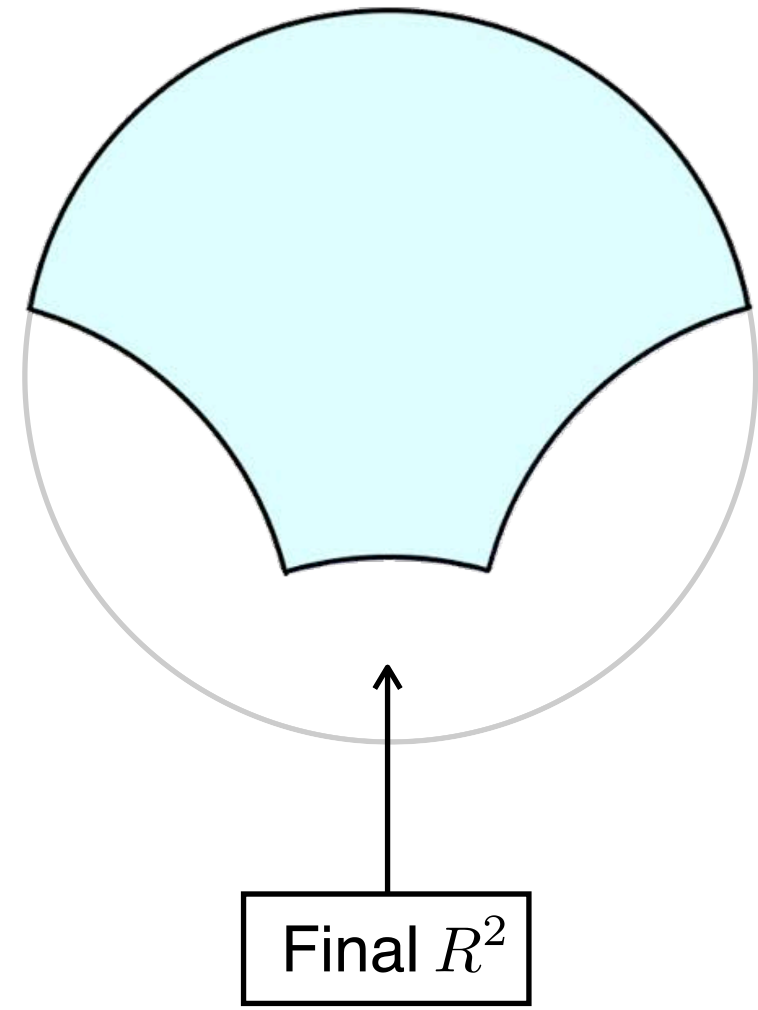


Adding a further predictor: X_3



Final explained and unexplained parts of Y





Change in R^2

SPSS commands for sequential regression

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI R ANOVA CHANGE
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT ltimedrs
  /METHOD=ENTER lphyheal sstress
  /METHOD=ENTER menheal.
```

SPSS commands for sequential regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	SSTRESS, LPHYHEAL[a]	.	Enter
2	Mental health symptoms[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: LTIMEDRS

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.614[a]	.376530	.374	.328594	.376530	139.507	2	462	.000
2	.614[b]	.376778	.373	.328885	.000248	.183	1	461	.669

a. Predictors: (Constant), SSTRESS, LPHYHEAL

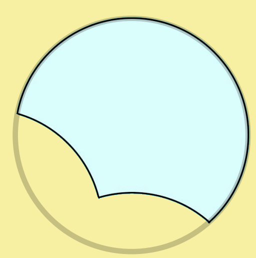
b. Predictors: (Constant), SSTRESS, LPHYHEAL, Mental health symptoms

SPSS commands for sequential regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	SSTRESS, LPHYHEAL[a]	.	Enter
2	Mental	.	Enter

This
is the initial R^2 value
for the model including
sstress and lphyheal:



Model	R	R Square	Square	Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.614[a]	.376530	.374	.328594	.376530	139.507	2	462	.000
2	.614[b]	.376778	.373	.328885	.000248	.183	1	461	.669

a. Predictors: (Constant), SSTRESS, LPHYHEAL

b. Predictors: (Constant), SSTRESS, LPHYHEAL, Mental health symptoms

SPSS commands for sequential regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	SSTRESS, LPHYHEAL[a]	.	Enter
2	Mental health [a]	.	Enter

a ... red.

Model	R	R Squ...			Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.614[a]	.37[illegible]30		[illegible]28594	.376530	139.507	2	462	.000
2	.614[b]	$\boxed{.376778}$	.373	.328885	.000248	.183	1	461	.669

a. Predictors: (Constant), SSTRESS, LPHYHEAL

b. Predictors: (Constant), SSTRESS, LPHYHEAL, Mental health symptoms

SPSS commands for sequential regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	SSTRESS, LPHYHEAL[a]	.	Enter
2	Mental health[a]	.	Enter

red.

This
is the difference in R^2
values between the two
models. That is, the *change* in R^2.

					Change Statistics				
Model	R	R Squar			R Square hange	F Change	df1	df2	Sig. F Change
1	.614[a]	.376530		28594	76530	139.507	2	462	.000
2	.614[b]	.376778	.373	.328885	.000248	.183	1	461	.669

a. Predictors: (Constant), SSTRESS, LPHYHEAL

b. Predictors: (Constant), SSTRESS, LPHYHEAL, Mental health symptoms

SPSS commands for sequential regression

If the Sig F Change value is large (i.e., >.05), then we retain the null hypothesis that the inclusion of the mental health variable, indeed, does not result in a significant increase in R^2. ✔

If the Sig F Change value is small (i.e., <.05), then we reject the null hypothesis that the mental health variable does in fact result in a significant increase in R^2. ✘

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1		.	Enter
2	Mental health symptoms[a]	.	Enter

a. All requ

b. Dep

This is the associated F value, degrees of freedom, and p value for this difference in R^2 values.

Model	R	R Square	Adju Squa		Change Statistics				
						F Change	df1	df2	Sig. F Change
1	.614[a]	.376530	.374		.37653	139.507	2	462	.000
2	.614[b]	.376778	.373	.328885	.000248	.183	1	461	.669

a. Predictors: (Constant), SSTRESS, LPHYHEAL

b. Predictors: (Constant), SSTRESS, LPHYHEAL, Mental health symptoms

Conclusion

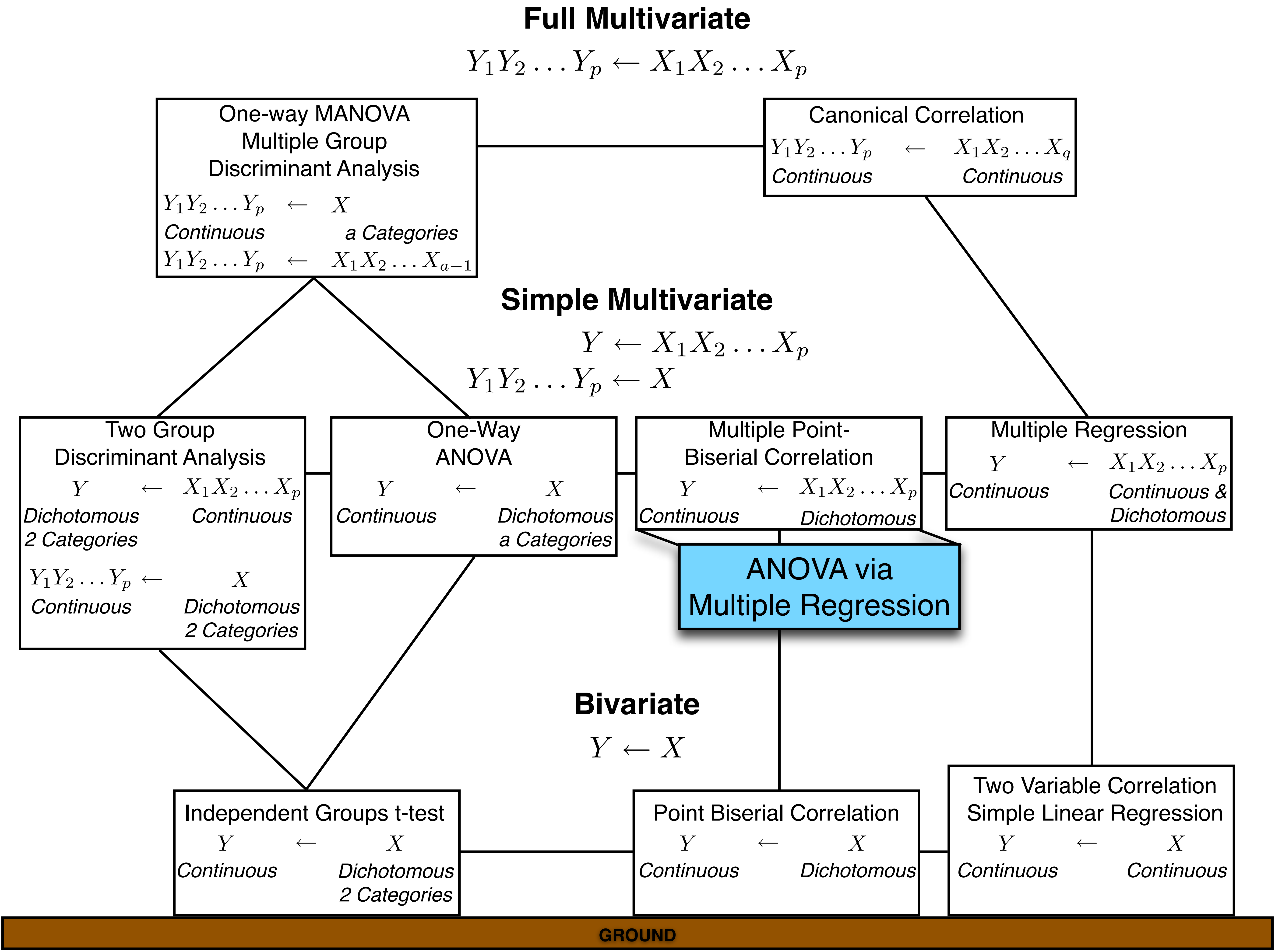
The combination of physical health and stress usefully predicts the number of visits to health professionals.

After differences in physical health and stress are controlled for, mental health does *not* add to the prediction of number of visits to health professionals.

The answer to the Tabachnick and Fidell's research question is that information regarding mental health after differences in physical health and stress are controlled for does *not* add to the prediction of number of visits to health professionals.

ANOVA via multiple regression

- ANOVA as a link in the Family Tree.
- Overview.
- t-test as a correlation
- Dummy coding of categorical variables
- Example
- Use of categorical variables in multiple regression


Full Multivariate
$Y_1 Y_2 \dots Y_p \leftarrow X_1 X_2 \dots X_p$
One-way MANOVA
Multiple Group
Discriminant Analysis
$Y_1 Y_2 \dots Y_p \leftarrow X$
Continuous
a Categories
$Y_1 Y_2 \dots Y_p \leftarrow X_1 X_2 \dots X_{a-1}$
Canonical Correlation
$Y_1 Y_2 \dots Y_p \leftarrow X_1 X_2 \dots X_q$
Continuous
Continuous
Simple Multivariate
$Y \leftarrow X_1 X_2 \dots X_p$
$Y_1 Y_2 \dots Y_p \leftarrow X$
Two Group
Discriminant Analysis
$Y \leftarrow X_1 X_2 \dots X_p$
Dichotomous
2 Categories
Continuous
$Y_1 Y_2 \dots Y_p \leftarrow X$
Continuous
Dichotomous
2 Categories
One-Way
ANOVA
$Y \leftarrow X$
Continuous
Dichotomous
a Categories
Multiple Point-
Biserial Correlation
$Y \leftarrow X_1 X_2 \dots X_p$
Continuous
Dichotomous
Multiple Regression
$Y \leftarrow X_1 X_2 \dots X_p$
Continuous
Continuous &
Dichotomous
ANOVA via
Multiple Regression
Bivariate
$Y \leftarrow X$
Independent Groups t-test
$Y \leftarrow X$
Continuous
Dichotomous
2 Categories
Point Biserial Correlation
$Y \leftarrow X$
Continuous
Dichotomous
Two Variable Correlation
Simple Linear Regression
$Y \leftarrow X$
Continuous
Continuous
GROUND

ANOVA via multiple regression: Overview

- Predictors (IVs) for regression: continuous or dichotomous
- The Link (The Trick) – IV for ANOVA: categorical (2+ levels)
- Convert a categorical variable into multiple dichotomous variables, then do an ANOVA using multiple regression – use linear composites.

t-test: Usual representation

Scores on Dependent Variable (Y)

Group 1	Group 2
10	8
9	7
11	9
11	8
10	5
$Y_1 = 10.2$	$Y_2 = 7.4$

$t = 3.61$

t-test: Alternative representation

Scores on Dependent Variable (Y)	Group Code (X)
10	1
9	1
11	1
11	1
10	1
8	2
7	2
9	2
8	2
5	2

Note: Dummy coding used for X.
If case is in Group 1 then X = 1
If case is in Group 2 then X = 2
$r_{Y,X} = -.7876$

Multiple regression:

$Y_{(cont)} \leftarrow X_1, X_2, \ldots X_p$ (all continuous or dichotomous)

ANOVA:

$Y_{(cont)} \leftarrow X_{(\text{categorical variable: a levels})}$

ANOVA by multiple regression:

$Y_{(cont)} \leftarrow X_1, X_2, \ldots X_{a-1}$ (a-1 dichotomous variables)

Types of Coding

- One of the tricks is in the coding of the categorical variable into dichotomous variables.
  - dummy coding
  - effect coding
  - orthogonal coding

ANOVA data: Usual representation

Data from three groups

A_1	A_2	A_3
4	7	1
5	8	2
6	9	3
7	10	4
8	11	5
$\overline{A}_1 = 6$	$\overline{A}_2 = 9$	$\overline{A}_3 = 3$

ANOVA data: Alternative representation

		Dummy Coding	
	Y	X_{D1}	X_{D2}
A_1	4	1	0
	5	1	0
	6	1	0
	7	1	0
	8	1	0
A_2	7	0	1
	8	0	1
	9	0	1
	10	0	1
	11	0	1
A_3	1	0	0
	2	0	0
	3	0	0
	4	0	0
	5	0	0

$Y \leftarrow X_{D1} \quad X_{D2}$

Tests of significance

- In ANOVA:
  - F ratio for the equality of means with (a-1) and [N-(a-1)-1] df

ANOVA

Y

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	90.000	2	45.000	18.000	.000
Within Groups	30.000	12	2.500		
Total	120.000	14			

- In regression:
  - F ratio for R^2 for the full model with (a-1) and N-p-1 df

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.866[a]	.750	.708	1.58114

a. Predictors: (Constant), D2, D1

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	90.000	2	45.000	18.000	.000[a]
	Residual	30.000	12	2.500		
	Total	120.000	14			

a. Predictors: (Constant), D2, D1

b. Dependent Variable: Y

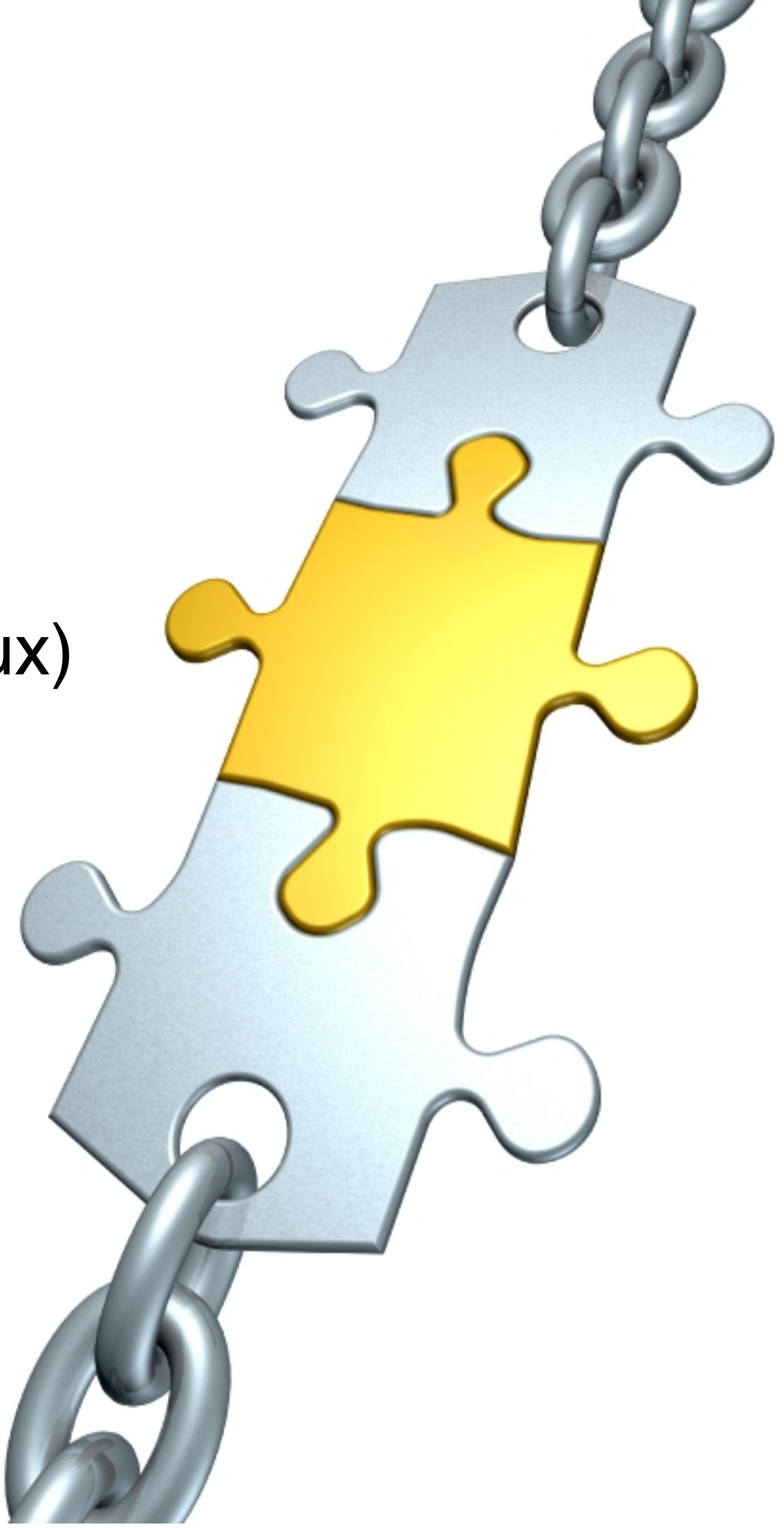
Summary

Being able to include dichotomous variables (which can represent categorical variables) opens the way to a much broader role for multiple regression.

In terms of the family tree of multivariate methods, using multiple regression to do an ANOVA provides the ‘missing link’ between the correlational and analysis of variance methods.

Moderated Multiple Regression

- A motivational example
- The case of the third variable (redux)
- Mediated or moderated?
- Interactions in ANOVA
- Interactions in multiple regression.



Moderated Multiple Regression

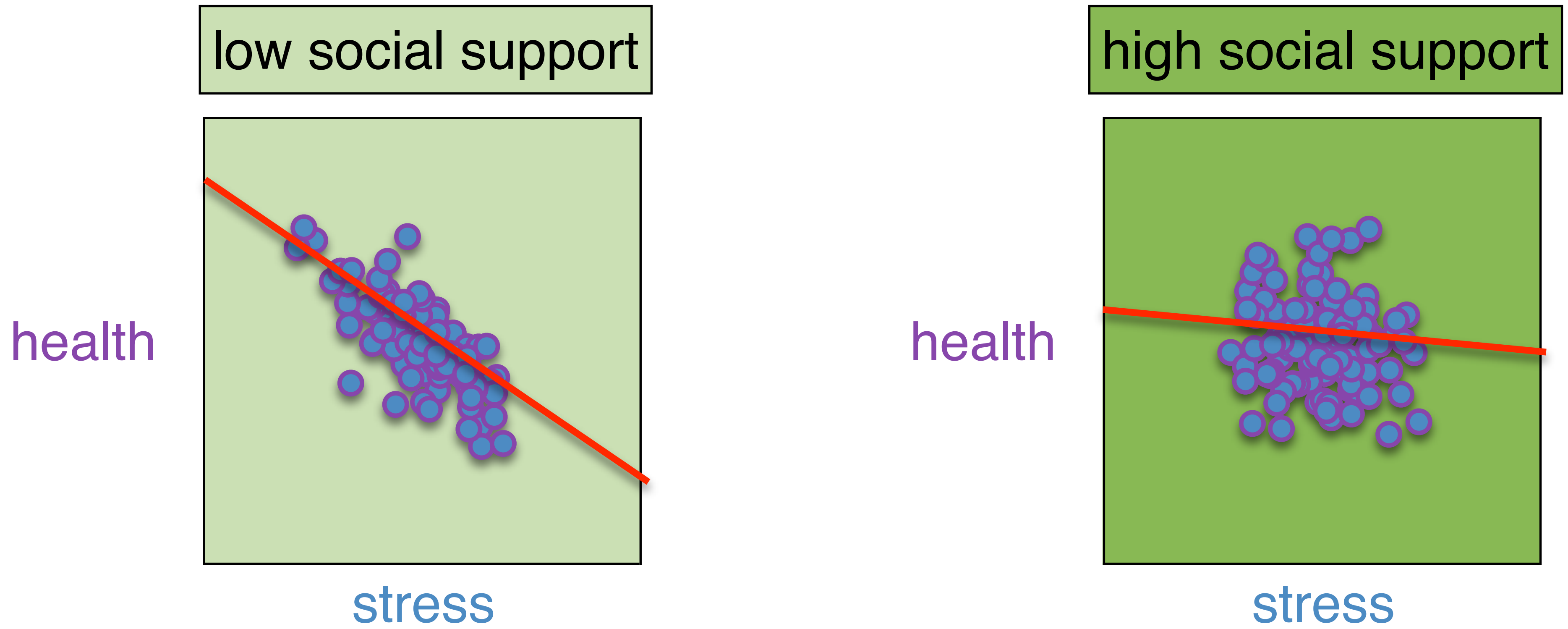
- A motivational example
  - Assume you are part of a team of researchers interested in health and predictors of it, especially stress.
  - Your research group thinks that social support also has an influence. From your own experiences and the literature, you suggest the following relationships among the variables:

When social support is low, there is a strong relationship between health and stress with high levels of stress leading to low health outcomes.

When social support is high, there is a very weak relationship between health and stress.

Moderated Multiple Regression

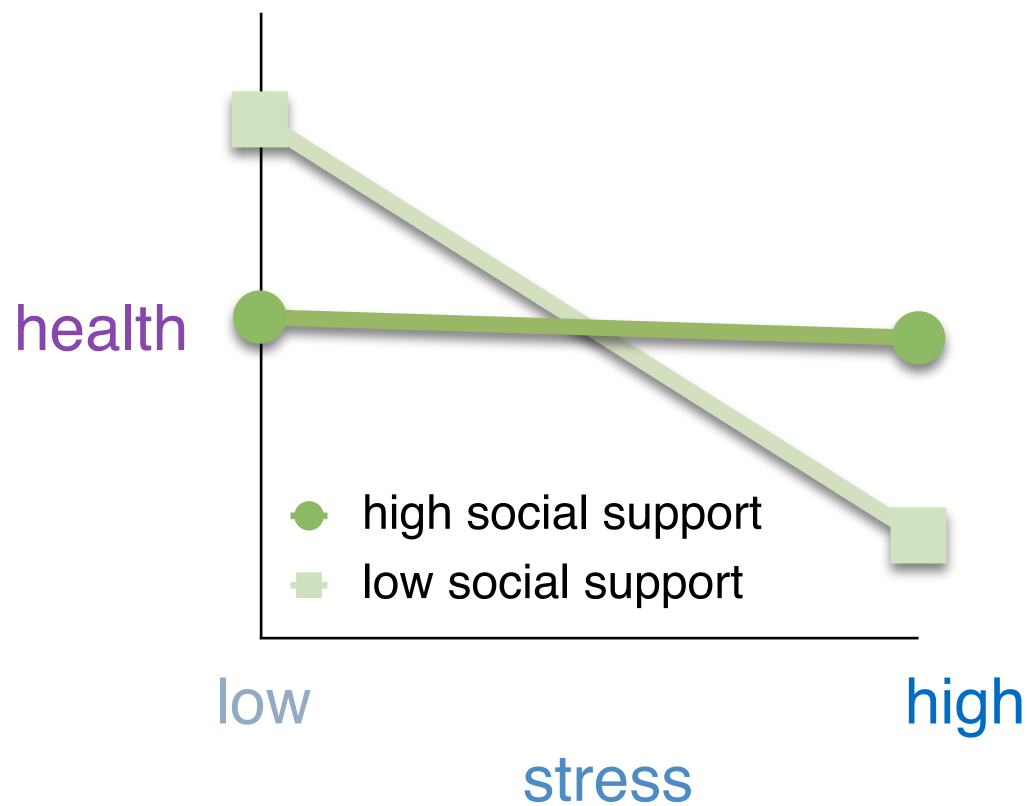
- A motivational example
  - Displaying the relationship between three variables (scatterplot view).



  - You think in terms of the variables being continuous.

Moderated Multiple Regression

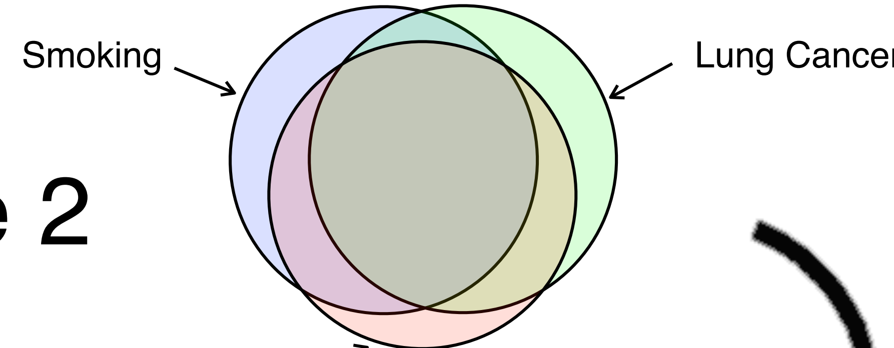
- A motivational example
  - ANOVA between groups view: “Some people only know ANOVA designs”:



  - Interpretation: There is an effect of stress only for the low social support group.

The Case of the Third Variable

Lecture 2



- Stress
- Coffee
- Gender

The Case of the Third Variable:

Some adjectives…

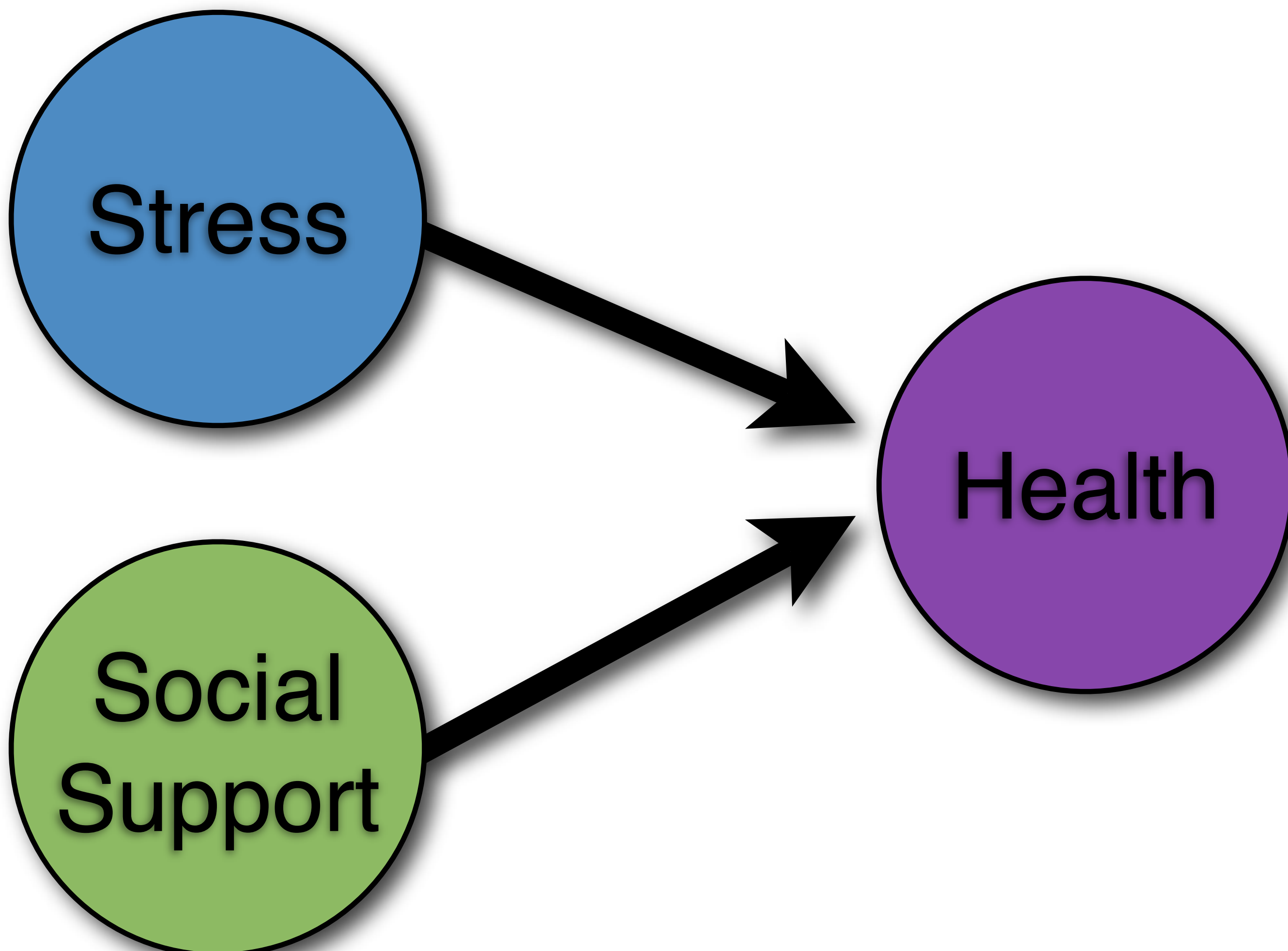
- Annoying is one
- Nuisance is another
- Confounding
- Extraneous
- Mediating
- Moderating

Your research group seems to be mixed up about these last two adjectives. So you sort them out...

The Case of the Third Variable:

Some adjectives…

Additive Effects: Main effects only



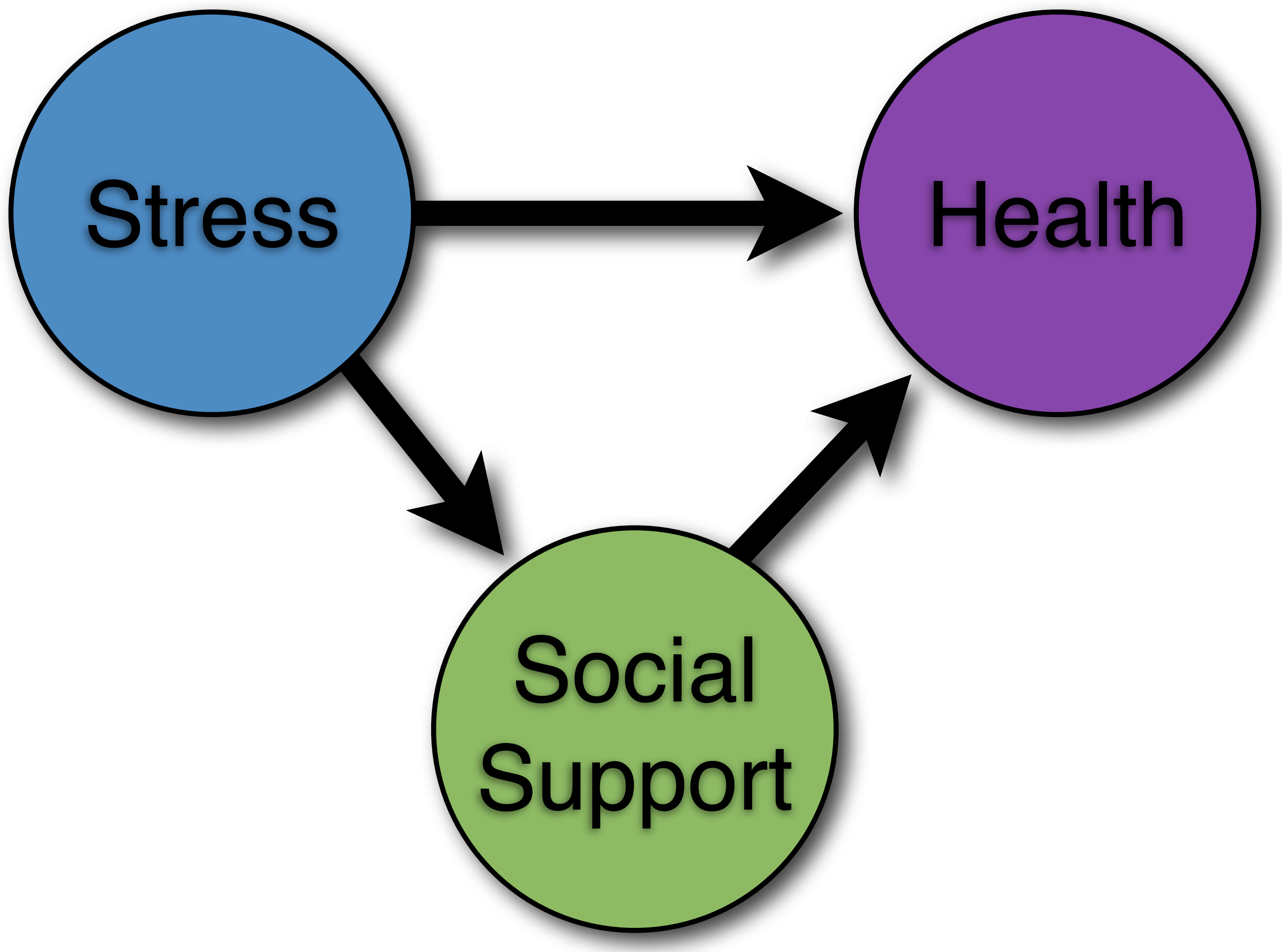
Schematic:

$$Y \leftarrow X_1 \; X_2$$

The Case of the Third Variable:

Some adjectives…

Mediated effects:
Stress affects health directly and indirectly through social support.

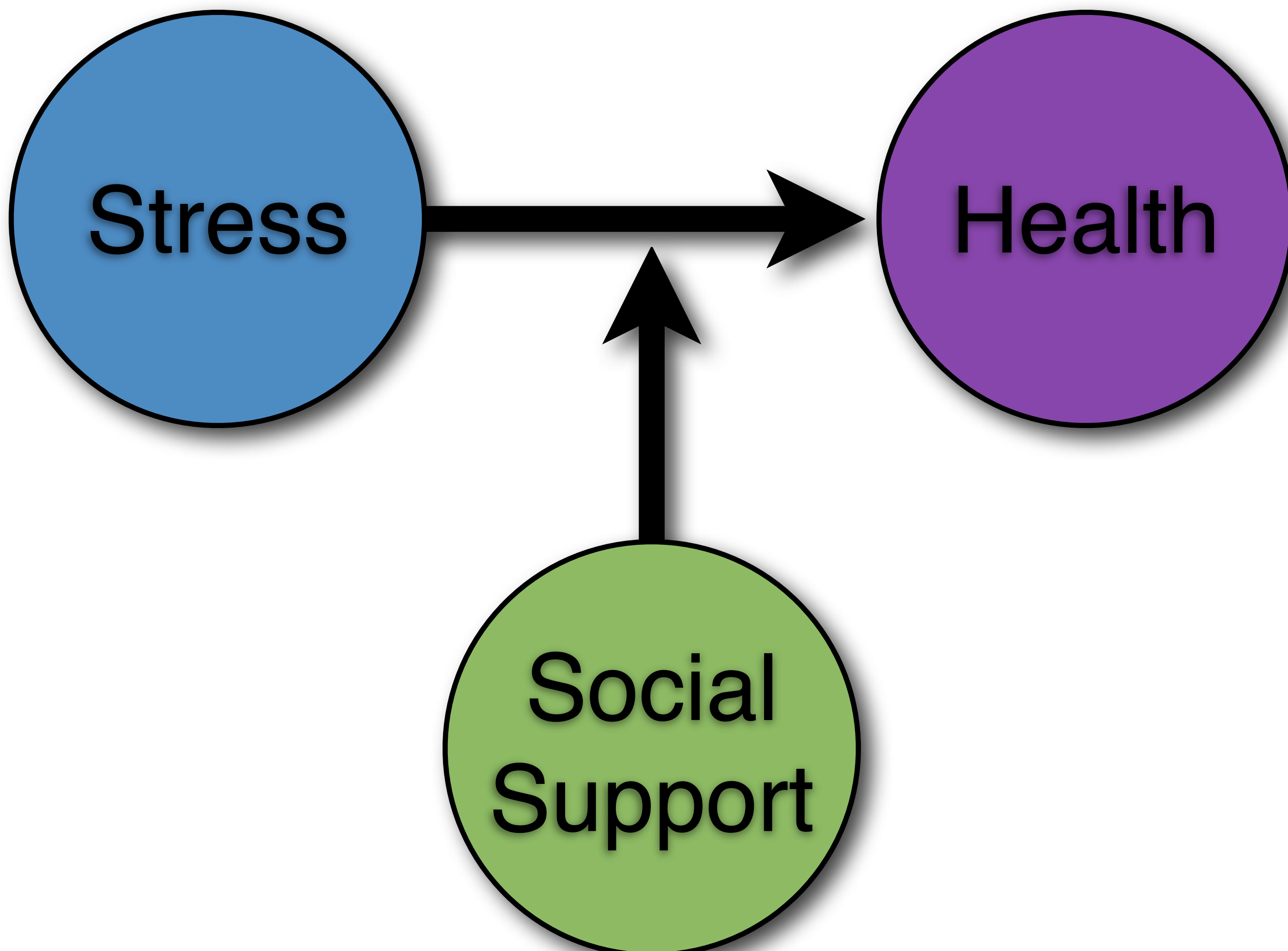


No easy schematic representation

The Case of the Third Variable:

Some adjectives...

Moderated effects:
Relationship between stress and health affected by social support.



Schematic representation later...

Slope of the regression: health $\leftarrow$ stress is different for different values of social support.

How to research moderated effects

- By extending a one-way ANOVA to a two-way ANOVA, we include another independent variable.
- The interesting part of a two-way ANOVA design is its ability to test for interactions.
  - This tests whether the effects of A on Y is moderated by B.

		social support	
		high	low
stress	high		
	low		

Three effects are tested in a two-way ANOVA:

The stress main effect, the social support main effect and the stress x social support interaction.

2x2 ANOVA by multiple regression

- Health is continuous. Stress and social support are both dichotomous in a two-way ANOVA. A multiple regression approach allows stress and social support to remain continuous (more information).
- Additive model:
  - health ← stress, social support
- Moderated effects model:
  - health ← stress, social support, stress x social support
- A score on the stress x social support variable is the product of stress and social support scores.

Whether the interaction contributes to the prediction of health is tested by the change in the amount of variance accounted for when the interaction term is added to the regression equation. This is done using a sequential regression strategy. The main effects are included in the model first, and then the interaction term is included. The R^2 change and its test for statistical significance indicates whether the interaction is significant.

Displaying the interaction with continuous predictors

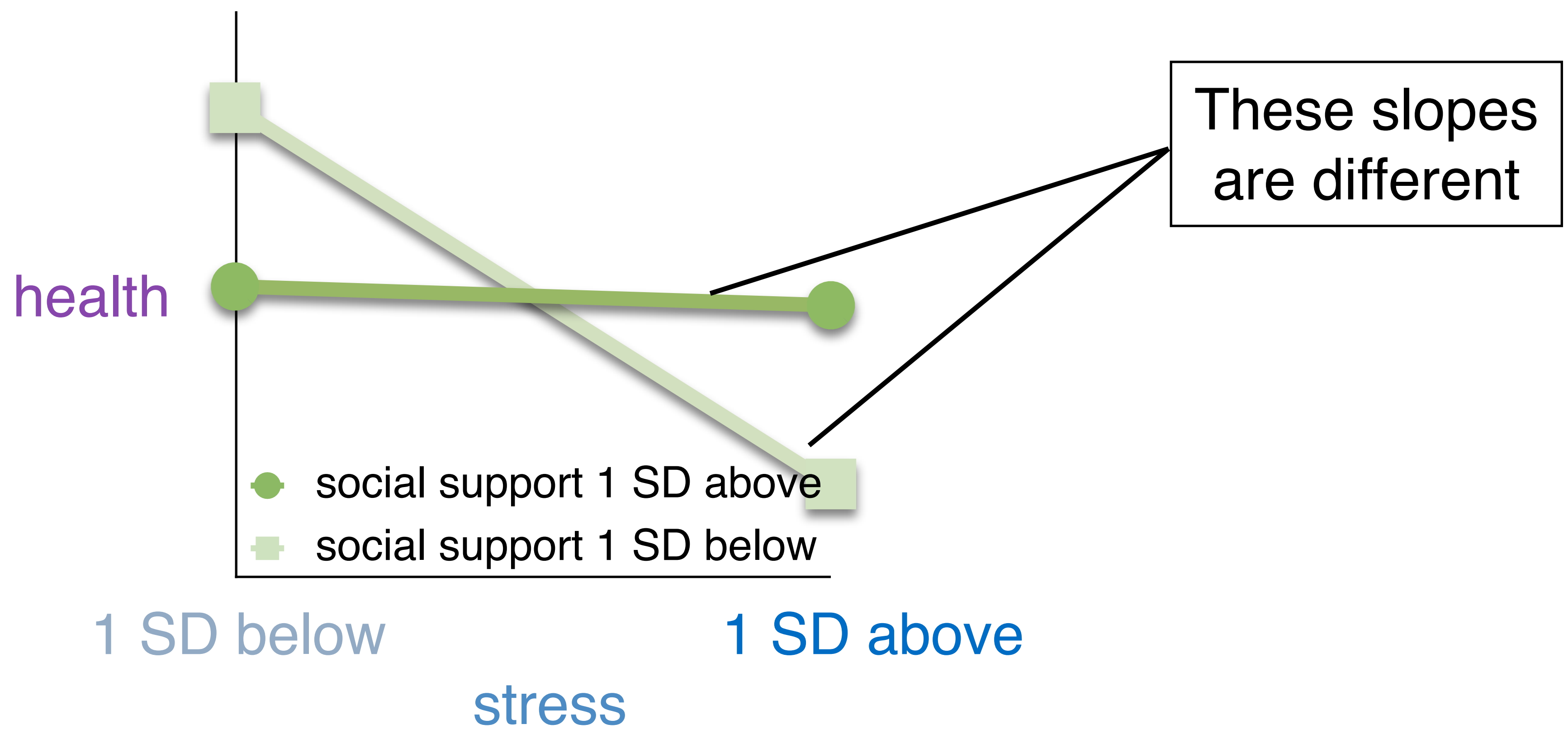
- Post-hoc tests are needed to verify that the pattern of relationships is what you expected them to be.
- A rough check using your scatterplot view will show the relationships in the expected directions.
- But you need something better than the eyeball test.
- A 'simple slopes' analysis:
  - What are the slopes of the relationship between health and stress at different levels of social support?
  - The model is known:

    $$Y' = a + b_1(\text{stress}) + b_2(\text{social support}) + b_3(\text{stress} \times \text{social support})$$

  - Calculate the predicted values of health given this model.

Displaying the interaction with continuous predictors

- A 'simple slopes' analysis:
  - Pick the points at one standard deviation from the mean on both stress and social support.
  - Clever use of the regression equations and using SPSS as the calculator gives the information to plot and test the simple slopes for statistical significance.
  - This is analogous to post-hoc 'simple main effects' tests in two-way ANOVA designs.



Summary

- The great advance from one-way ANOVA to two-way designs is the ability to test for the interaction effect.
- Multiple regression allows these interaction effects to be tested when variables are continuous.
- Multiple regression is a very flexible data analytic tool.