Psychology 516
Applied Multivariate Analysis

Major Goals:

- Expand your repertoire of analytical options.
- Understand enough theory to appreciate appropriate and inappropriate application.
- Be able to apply the methods using R.
- Know where to go for additional help.

Today's Goal:

Understand the central role that linear combinations play in nearly all statistical procedures.

The basic starting point for any statistical analysis is a *matrix* of data. For most applications in the social sciences, this matrix will be a People x Variables array.

But, the *objects of measurement* need not be people—they could be animals, work groups, cities, etc.

"The numbers do not know where they came from." *F. Lord*

	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}	. . .	V_K
P_1												
P_2												
P_3												
P_4												
P_5												
. . .												
P_N												

The variables (V) can be continuous measures, categories represented by numbers, transformations, products or combinations of other variables.

Nearly all statistical procedures—univariate and multivariate—are based on *linear combinations*. Understanding that basic fact has far-reaching implications for using statistical procedures to their fullest advantage.

A linear combination (LC) for a particular person (i) is nothing more than a weighted (W) sum of variables (V):

$$LC_i = W_1V_{i,1} + W_2V_{i,2} + . . . + W_KV_{i,K}$$

$LC_i = W_1V_{i,1} + W_2V_{i,2} + \ldots + W_KV_{i,K}$

A very simple example is the total score on a questionnaire. The individual items on the questionnaire are the variables V_1, V_2, V_3, etc. The weights are all set to a value of 1 (i.e., $W_1 = W_2 = \ldots W_k = 1$).

What assumption underlies this linear combination? *Why* do we combine the variables in this way?

The items combined in a linear combination need not be variables. In statistics, the items combined are often people (P).

$LC_j = W_1P_{1,j} + W_2P_{2,j} + \ldots + W_NP_{N,j}$

A good example is the sample mean. In this case the weights are set to the reciprocal of the sample size (i.e., $W_1 = W_2 = \ldots W_k = 1/N$).

What assumption underlies this linear combination?

Another common form of linear combination in statistics is a weighted combination of means.

	Treatment	Control
Men	$M_{M,T}$	$M_{M,C}$
Women	$M_{W,T}$	$M_{W,C}$

What is the purpose of the following linear combination of the means?

$LC = (1)M_{M,T} + (-1)M_{M,C} + (1)M_{W,T} + (-1)M_{W,C}$

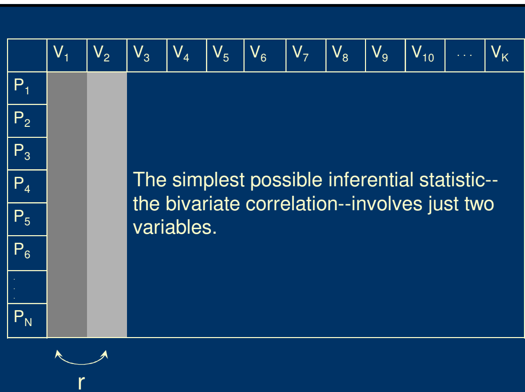
What weights would be necessary to know if treatment was more effective for men than for women?

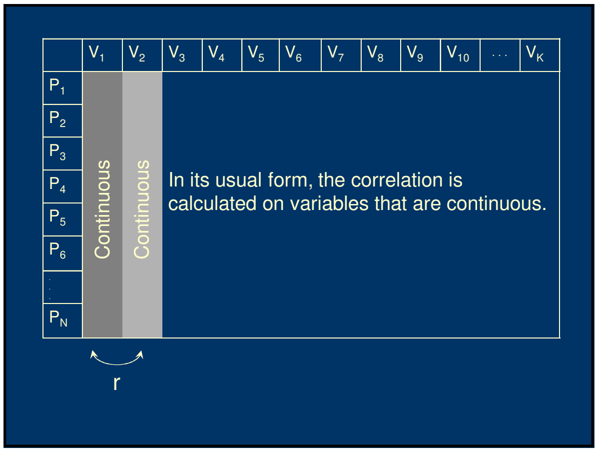
What assumption underlies these linear combinations?

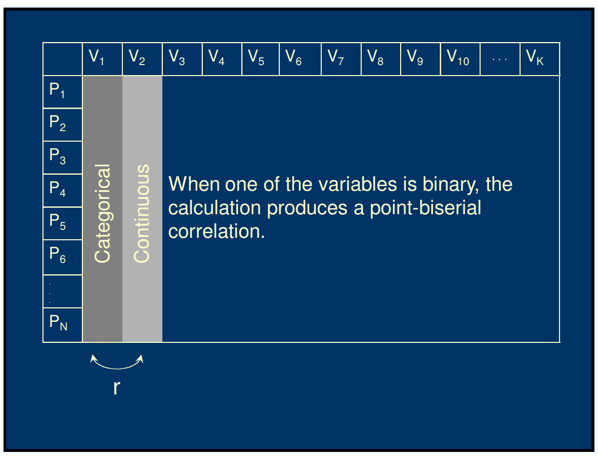
	Treatment	Control
Men	$M_{M,T}$	$M_{M,C}$
Women	$M_{W,T}$	$M_{W,C}$

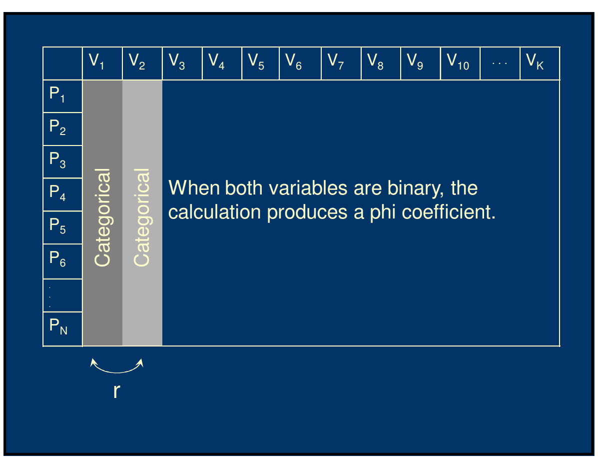
Different statistical procedures derive the weights (W) in a linear combination to either *maximize some desirable property* (e.g., a correlation or an effect) or to *minimize some undesirable property* (e.g., error).

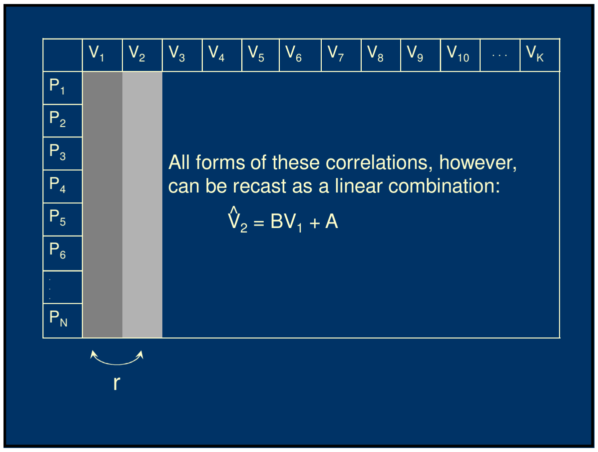
The weights are sometimes *empirically* determined and sometimes they are dictated by *theory* (e.g., dummy, effect, and contrast codes) to produce linear combinations of particular interest.

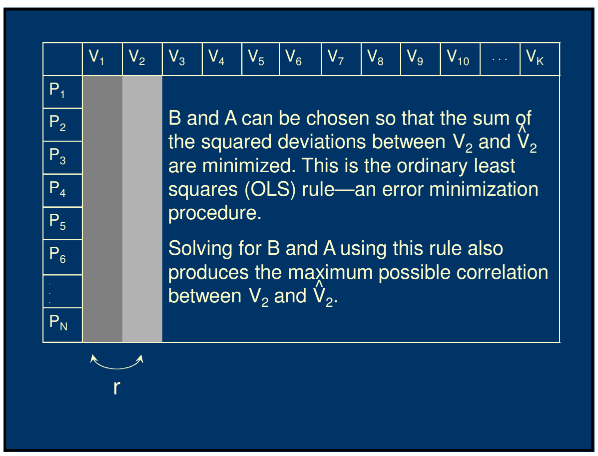


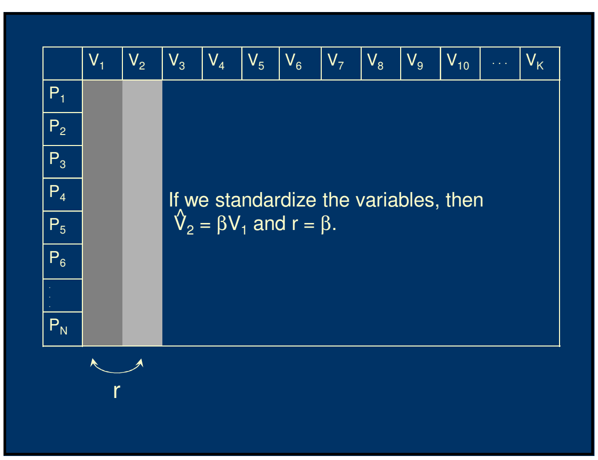

V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 . . . VK
P1 P2 P3 P4 P5 P6 . . . PN
Continuous
Continuous
In its usual form, the correlation is calculated on variables that are continuous.
r

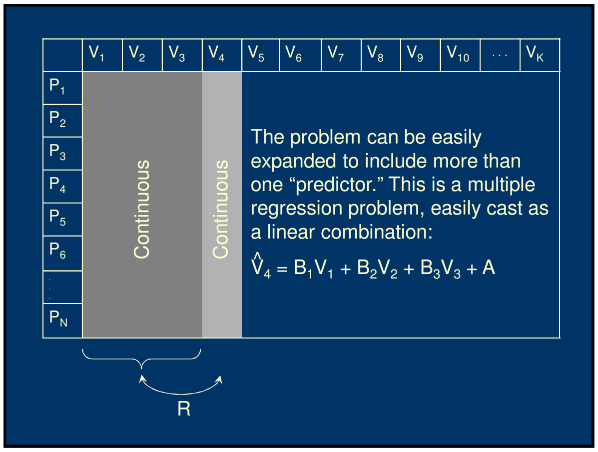

V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 . . . VK
P1 P2 P3 P4 P5 P6 . . . PN
Categorical
Continuous
When one of the variables is binary, the calculation produces a point-biserial correlation.
r


V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 . . . VK
P1 P2 P3 P4 P5 P6 . . . PN
Categorical
Categorical
When both variables are binary, the calculation produces a phi coefficient.
r

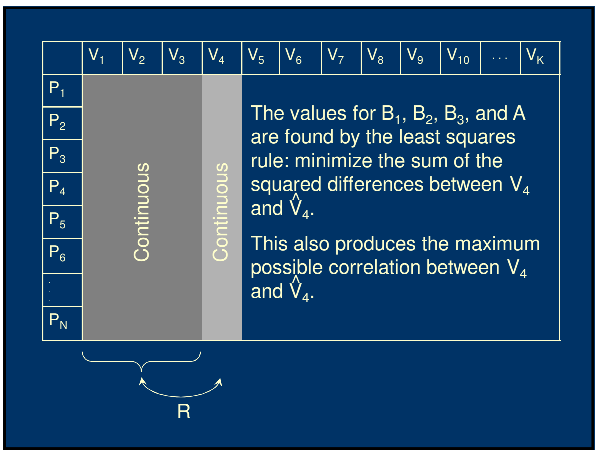

V_1 V_2 V_3 V_4 V_5 V_6 V_7 V_8 V_9 V_{10} … V_K
P_1 P_2 P_3 P_4 P_5 P_6 … P_N
All forms of these correlations, however, can be recast as a linear combination:
$\hat{V}_2 = BV_1 + A$
r


V_1 V_2 V_3 V_4 V_5 V_6 V_7 V_8 V_9 V_{10} … V_K
P_1 P_2 P_3 P_4 P_5 P_6 … P_N
B and A can be chosen so that the sum of the squared deviations between V_2 and $\hat{V}_2$ are minimized. This is the ordinary least squares (OLS) rule—an error minimization procedure.
Solving for B and A using this rule also produces the maximum possible correlation between V_2 and $\hat{V}_2$.
r

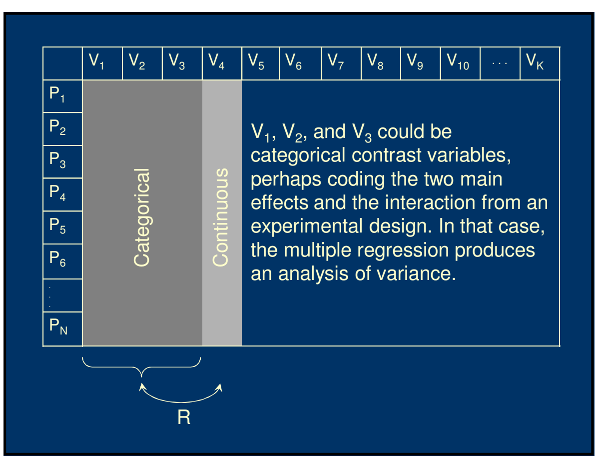

V_1 V_2 V_3 V_4 V_5 V_6 V_7 V_8 V_9 V_{10} … V_K
P_1 P_2 P_3 P_4 P_5 P_6 … P_N
If we standardize the variables, then $\hat{V}_2 = \beta V_1$ and $r = \beta$.
r



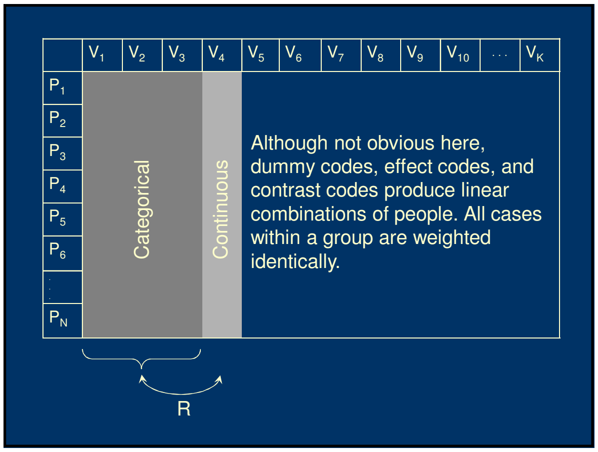
V_1 V_2 V_3 V_4 V_5 V_6 V_7 V_8 V_9 V_{10} ... V_K
P_1 P_2 P_3 P_4 P_5 P_6 P_N
Continuous
Continuous
The problem can be easily expanded to include more than one "predictor." This is a multiple regression problem, easily cast as a linear combination:
$\hat{V}_4 = B_1V_1 + B_2V_2 + B_3V_3 + A$
R



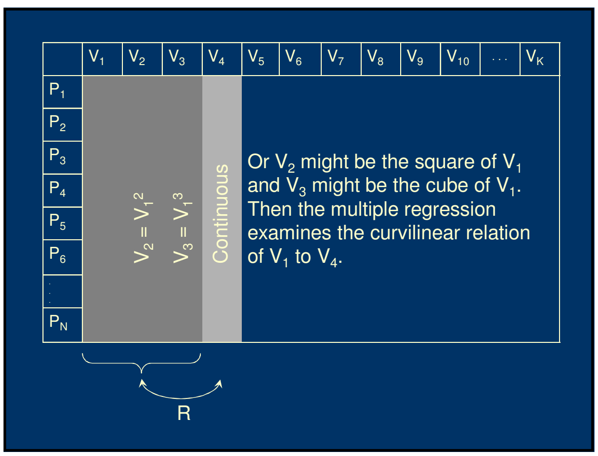
V_1 V_2 V_3 V_4 V_5 V_6 V_7 V_8 V_9 V_{10} ... V_K
P_1 P_2 P_3 P_4 P_5 P_6 P_N
Continuous
Continuous
The values for B_1, B_2, B_3, and A are found by the least squares rule: minimize the sum of the squared differences between V_4 and $\hat{V}_4$.
This also produces the maximum possible correlation between V_4 and $\hat{V}_4$.
R



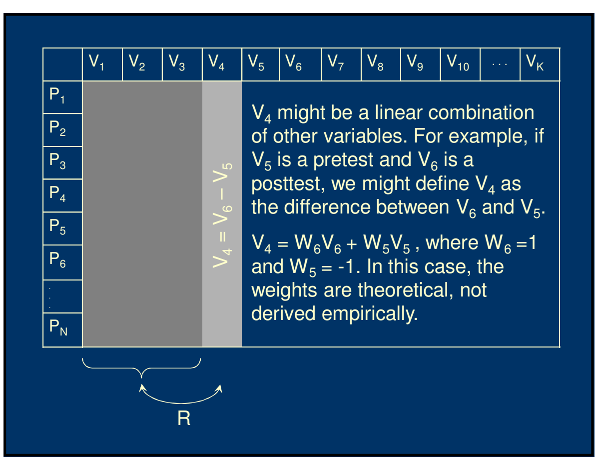
V_1 V_2 V_3 V_4 V_5 V_6 V_7 V_8 V_9 V_{10} ... V_K
P_1 P_2 P_3 P_4 P_5 P_6 P_N
Categorical
Continuous
V_1, V_2, and V_3 could be categorical contrast variables, perhaps coding the two main effects and the interaction from an experimental design. In that case, the multiple regression produces an analysis of variance.
R



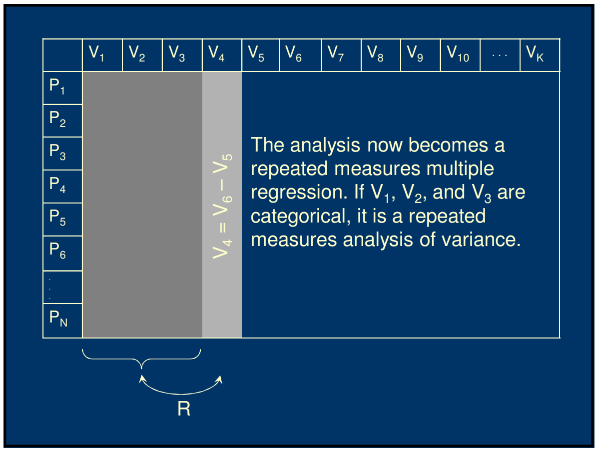
V_1 V_2 V_3 V_4 V_5 V_6 V_7 V_8 V_9 V_{10} ... V_K
P_1 P_2 P_3 P_4 P_5 P_6 . . . P_N
Categorical
Continuous
Although not obvious here, dummy codes, effect codes, and contrast codes produce linear combinations of people. All cases within a group are weighted identically.
R

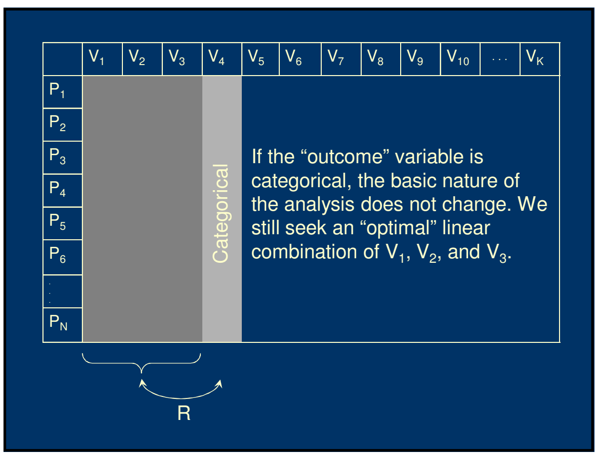


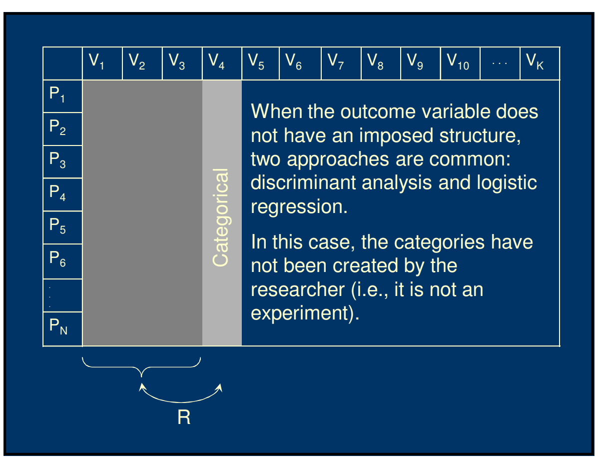
V_1 V_2 V_3 V_4 V_5 V_6 V_7 V_8 V_9 V_{10} ... V_K
P_1 P_2 P_3 P_4 P_5 P_6 . . . P_N
$V_2 = V_1^2$
$V_3 = V_1^3$
Continuous
Or V_2 might be the square of V_1 and V_3 might be the cube of V_1. Then the multiple regression examines the curvilinear relation of V_1 to V_4.
R

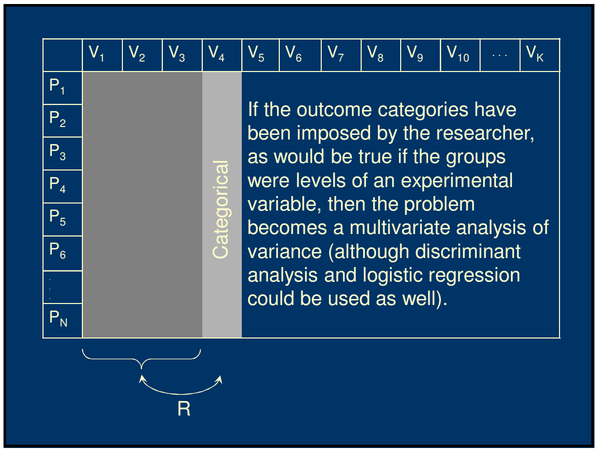


V_1 V_2 V_3 V_4 V_5 V_6 V_7 V_8 V_9 V_{10} ... V_K
P_1 P_2 P_3 P_4 P_5 P_6 . . . P_N
$V_4 = V_6 - V_5$
V_4 might be a linear combination of other variables. For example, if V_5 is a pretest and V_6 is a posttest, we might define V_4 as the difference between V_6 and V_5.
$V_4 = W_6V_6 + W_5V_5$, where $W_6 = 1$ and $W_5 = -1$. In this case, the weights are theoretical, not derived empirically.
R


V_1 V_2 V_3 V_4 V_5 V_6 V_7 V_8 V_9 V_{10} … V_K
P_1 P_2 P_3 P_4 P_5 P_6 … P_N
$V_4 = V_6 – V_5$
The analysis now becomes a repeated measures multiple regression. If V_1, V_2, and V_3 are categorical, it is a repeated measures analysis of variance.
R


V_1 V_2 V_3 V_4 V_5 V_6 V_7 V_8 V_9 V_{10} … V_K
P_1 P_2 P_3 P_4 P_5 P_6 … P_N
Categorical
If the “outcome” variable is categorical, the basic nature of the analysis does not change. We still seek an “optimal” linear combination of V_1, V_2, and V_3.
R


V_1 V_2 V_3 V_4 V_5 V_6 V_7 V_8 V_9 V_{10} … V_K
P_1 P_2 P_3 P_4 P_5 P_6 … P_N
Categorical
When the outcome variable does not have an imposed structure, two approaches are common: discriminant analysis and logistic regression.
In this case, the categories have not been created by the researcher (i.e., it is not an experiment).
R



produces the biggest mean difference between these two groups?

What linear combination of these variables . . .

	Men	Women
Variable 1		
Variable 2		
Variable 3		
Variable 4		
Variable 5		

best predicts membership in these two groups?

What linear combination of these variables . . .

	Men	Women
Variable 1		
Variable 2		
Variable 3		
Variable 4		
Variable 5		

What linear combination of these variables . . . produces the highest correlation with a binary variable coded 0 (men) and 1 (women)?

	Men	Women
Variable 1		
Variable 2		
Variable 3		
Variable 4		
Variable 5		

When labeled this way, it is easy to think of this as a *regression* problem.

"Predictors" "Outcome"

	Men	Women
Variable 1		
Variable 2		
Variable 3		
Variable 4		
Variable 5		

We can put any kind of group in as the "outcome." It doesn't change the nature of the analysis.

"Predictors" "Outcome"

	Treatment	Control
Variable 1		
Variable 2		
Variable 3		
Variable 4		
Variable 5		

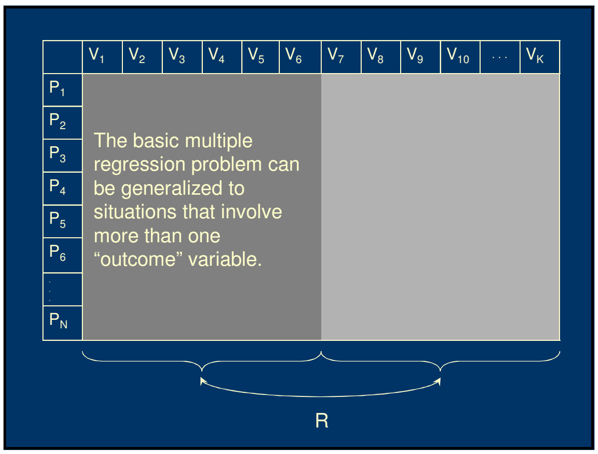
The more typical arrangement and labeling of experimental data would look like this and would be analyzed with analysis of variance . . .

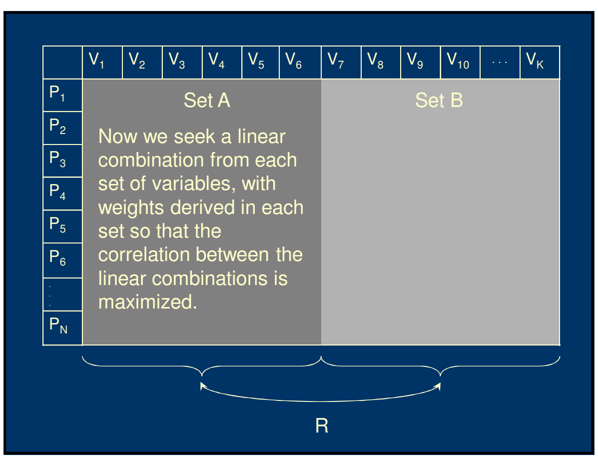
"Outcome"

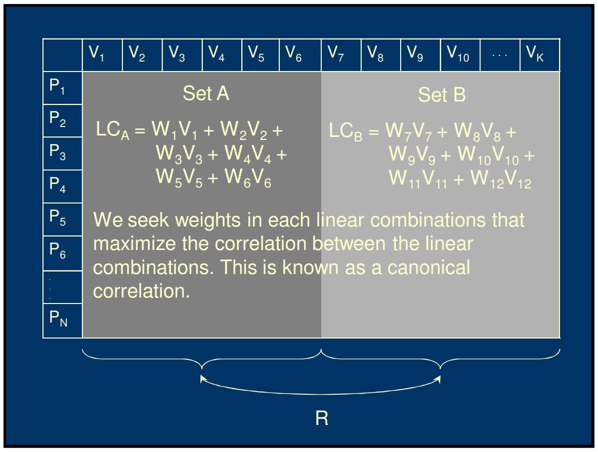
"Predictors"

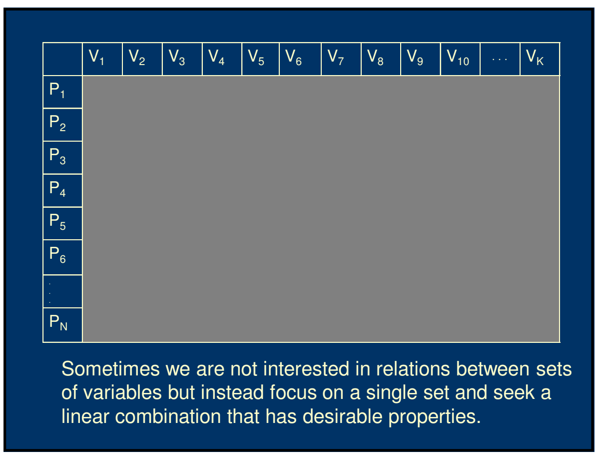
	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
Treatment					
Control					

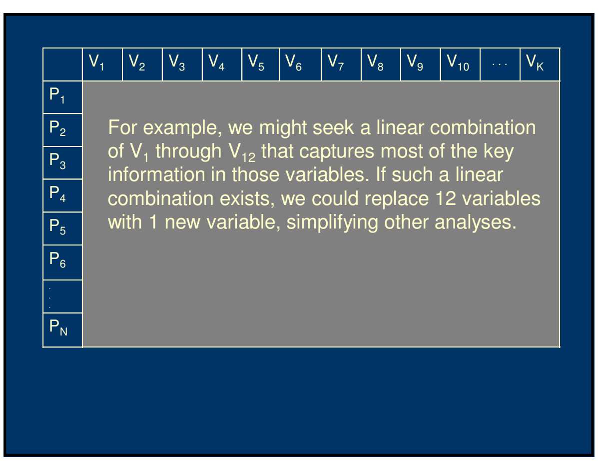
but a multivariate analysis of variance will still ask the question, "what linear combination of variables best separates the groups?"

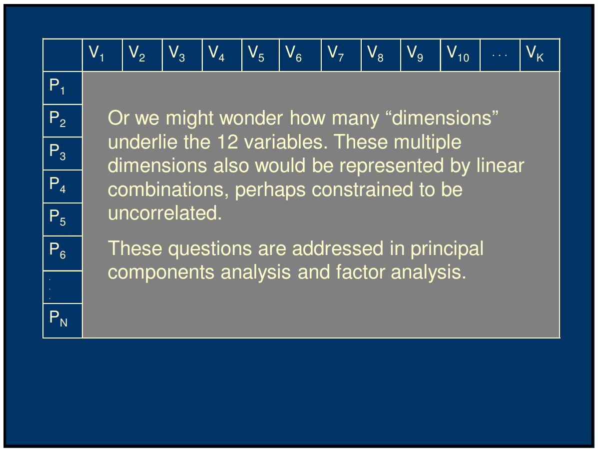




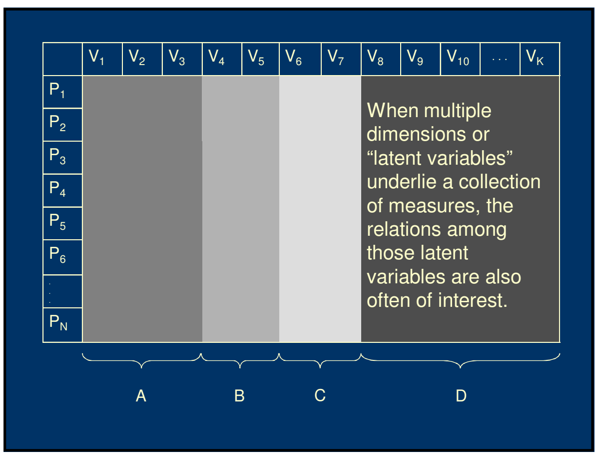

V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 ... VK
P1 P2 P3 P4 P5 P6 PN
Set A
$LC_A = W_1V_1 + W_2V_2 + W_3V_3 + W_4V_4 + W_5V_5 + W_6V_6$
Set B
$LC_B = W_7V_7 + W_8V_8 + W_9V_9 + W_{10}V_{10} + W_{11}V_{11} + W_{12}V_{12}$
We seek weights in each linear combinations that maximize the correlation between the linear combinations. This is known as a canonical correlation.
R


V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 ... VK
P1 P2 P3 P4 P5 P6 PN
Sometimes we are not interested in relations between sets of variables but instead focus on a single set and seek a linear combination that has desirable properties.

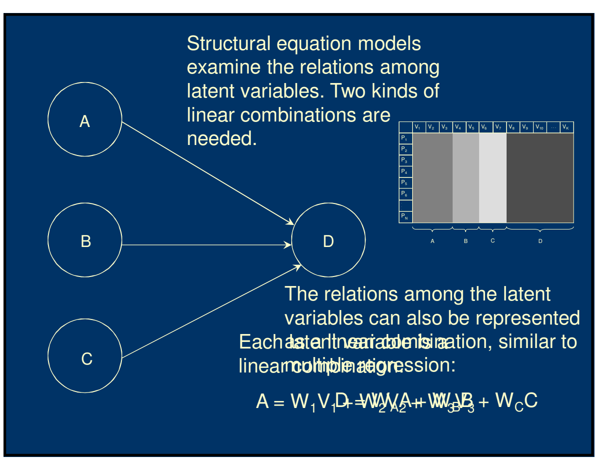

V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 ... VK
P1 P2 P3 P4 P5 P6 PN
For example, we might seek a linear combination of V_1 through V_{12} that captures most of the key information in those variables. If such a linear combination exists, we could replace 12 variables with 1 new variable, simplifying other analyses.



V_1 V_2 V_3 V_4 V_5 V_6 V_7 V_8 V_9 V_{10} . . . V_K
P_1 P_2 P_3 P_4 P_5 P_6 P_N
Or we might wonder how many “dimensions” underlie the 12 variables. These multiple dimensions also would be represented by linear combinations, perhaps constrained to be uncorrelated.
These questions are addressed in principal components analysis and factor analysis.



V_1 V_2 V_3 V_4 V_5 V_6 V_7 V_8 V_9 V_{10} . . . V_K
P_1 P_2 P_3 P_4 P_5 P_6 P_N
When multiple dimensions or “latent variables” underlie a collection of measures, the relations among those latent variables are also often of interest.
A
B
C
D



Structural equation models examine the relations among latent variables. Two kinds of linear combinations are needed.
A
B
C
D
Each latent variable is a linear combination:
$A = W_1V_1 + W_2V_2 + W_3V_3$
The relations among the latent variables can also be represented as multiple regression:
$D = W_AA + W_BB + W_CC$

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	. . .	P_N
V_1												
V_2												
V_3												
V_4												
V_5												
⋮												
V_K												

Sometimes we might shift the status of "people" and "variables" in our analysis. Our interest might be in whether a smaller number of dimensions or clusters might underlie the larger collection of people.

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	. . .	P_N
V_1												
V_2												
V_3												
V_4												
V_5												
⋮												
V_K												

Approaches such as multidimensional scaling and cluster analysis can address such questions. These are conceptually similar to principal components analysis, but on a *transposed matrix*.

The key idea is that the original data matrix can be transformed using linear combinations to provide useful ways to summarize the data and to test hypotheses about how the data are structured.

Sometimes the linear combinations are of variables and sometimes they are of people (or other useful objects of measurement). Sometimes both are of interest in the same analysis.

The goal of a statistical analysis is to get in close proximity to the truth. That requires thinking flexibly and creatively about the data and analyses.

"All models are wrong but some are useful."
G. E. P. Box

Next up . . . Matrix Algebra

Statistical formulas, especially multivariate formulations, are most conveniently expressed in matrix form and manipulated using matrix algebra.

Understanding basic matrix operations assists the optimal *construction* of linear combinations.