Readings

T&F: Chapters 1, 2, 3 and 5

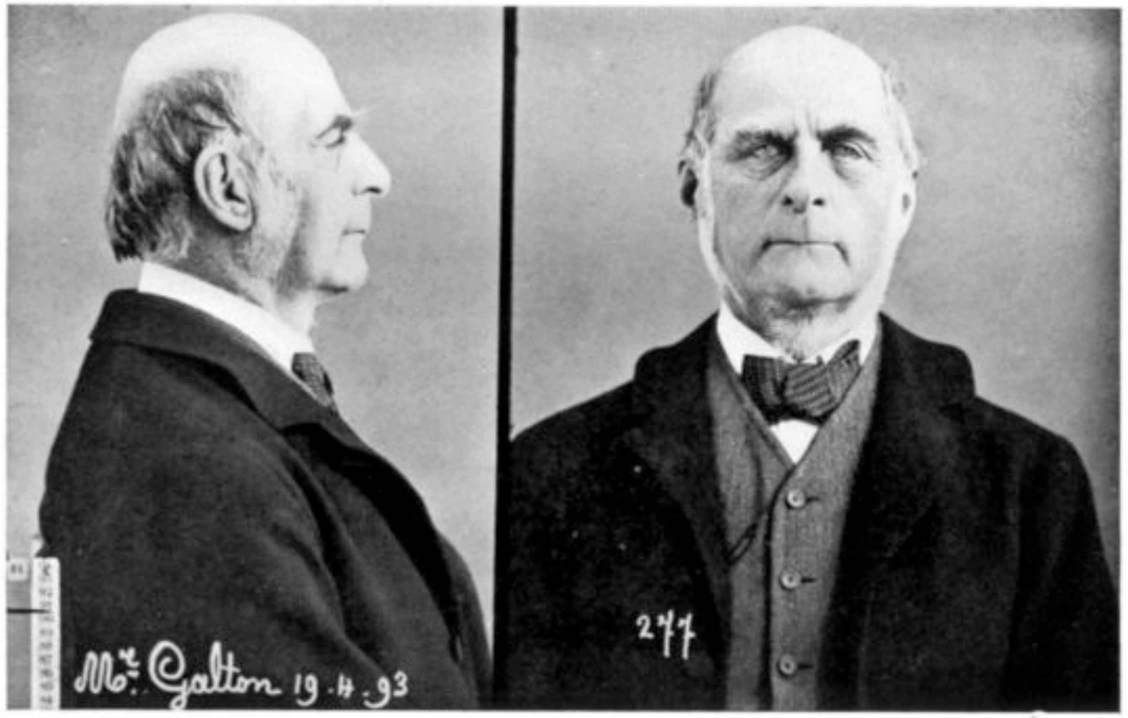
Linear Composites

A variable created by combining several existing variables.

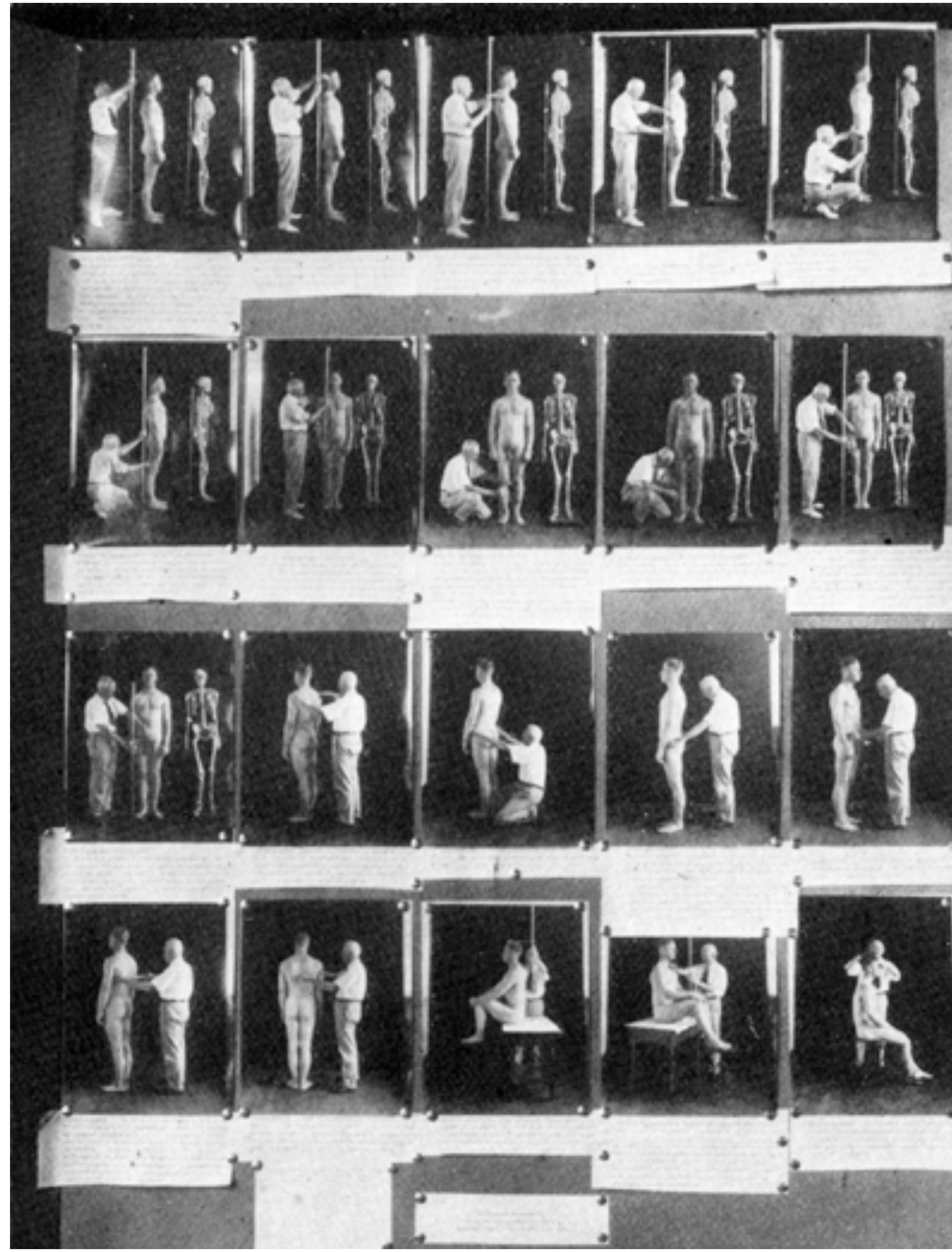
- each existing variable is given a weight.
- each variable is multiplied by its weight.
- weighted variables are added together to form a new variable.
- different weights will produce different linear composites.
- Example: Grade Point Average

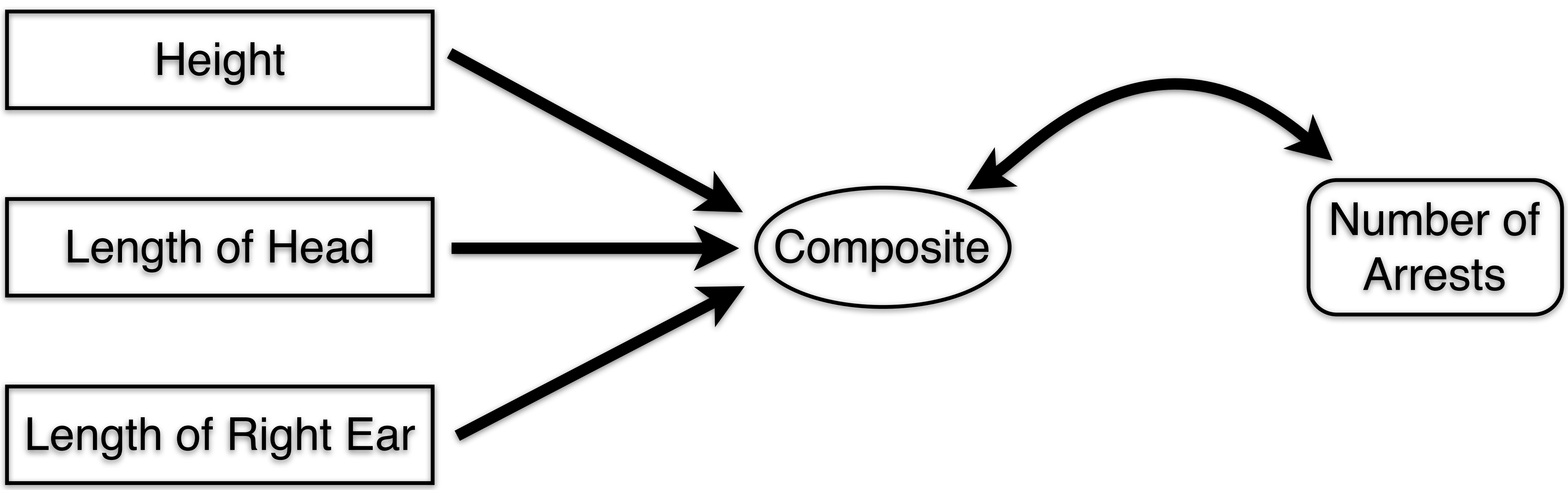
Taille 1m	Tête Long'	Pied g.	Coul' de l'iris g. N° de cl.	Agé de
Voûte	Tête Larg'	Médius g.	Aur'e	né le
Enverg 1m	Oreille dr. Long'	Auric'e g.	Pér'e	à
Buste 0,	Oreille dr. Larg'	Coudée g.	Part''	dép'
				Age app'

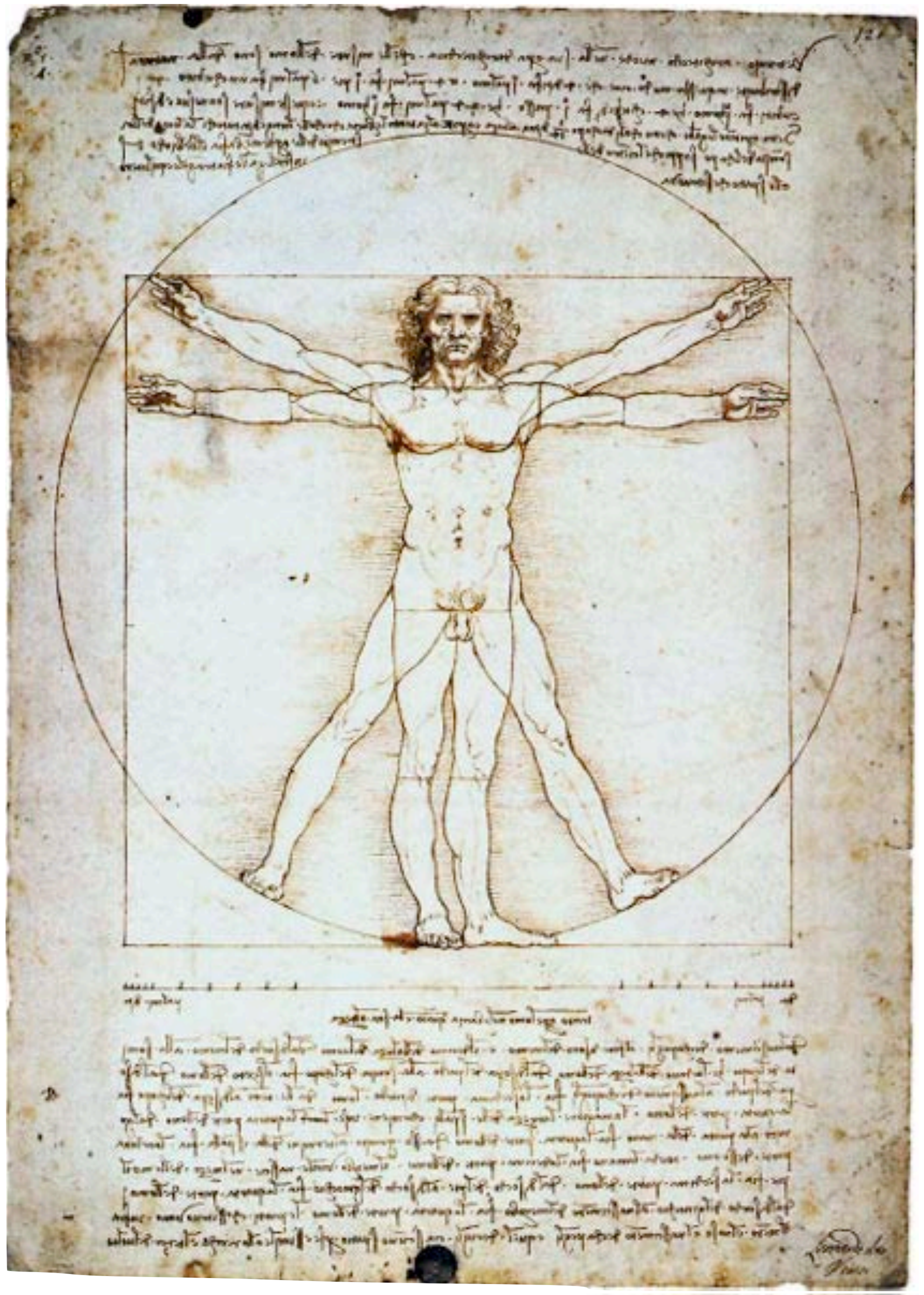
(Réduction photographique 1/7.)



Front. Inclin'	Nez. Racine (cavité)	Oreille droite. Bord o. s. p. f.	Barbe / Color' (pig''
Haut'	Dos Base	Lob. c. a. m. d.	Cheveux / sang'
Larg'	Haut' Saillie. Larg'	A. trg. i. p. r. d.	Car Ceint.
Part''	l l	Pli. f. s. h. E	*Autres traits caractéristiques :*
	Part''	Part.	Sig' dressé par M.




Height
Length of Head
Length of Right Ear
Composite
Number of
Arrests



Fluctuating Asymmetry (FA)

Composite = FA

“Results indicated that normally cycling (non-pill using) women near the peak fertility of their cycle tended to prefer the scent of shirts worn by symmetrical men.”

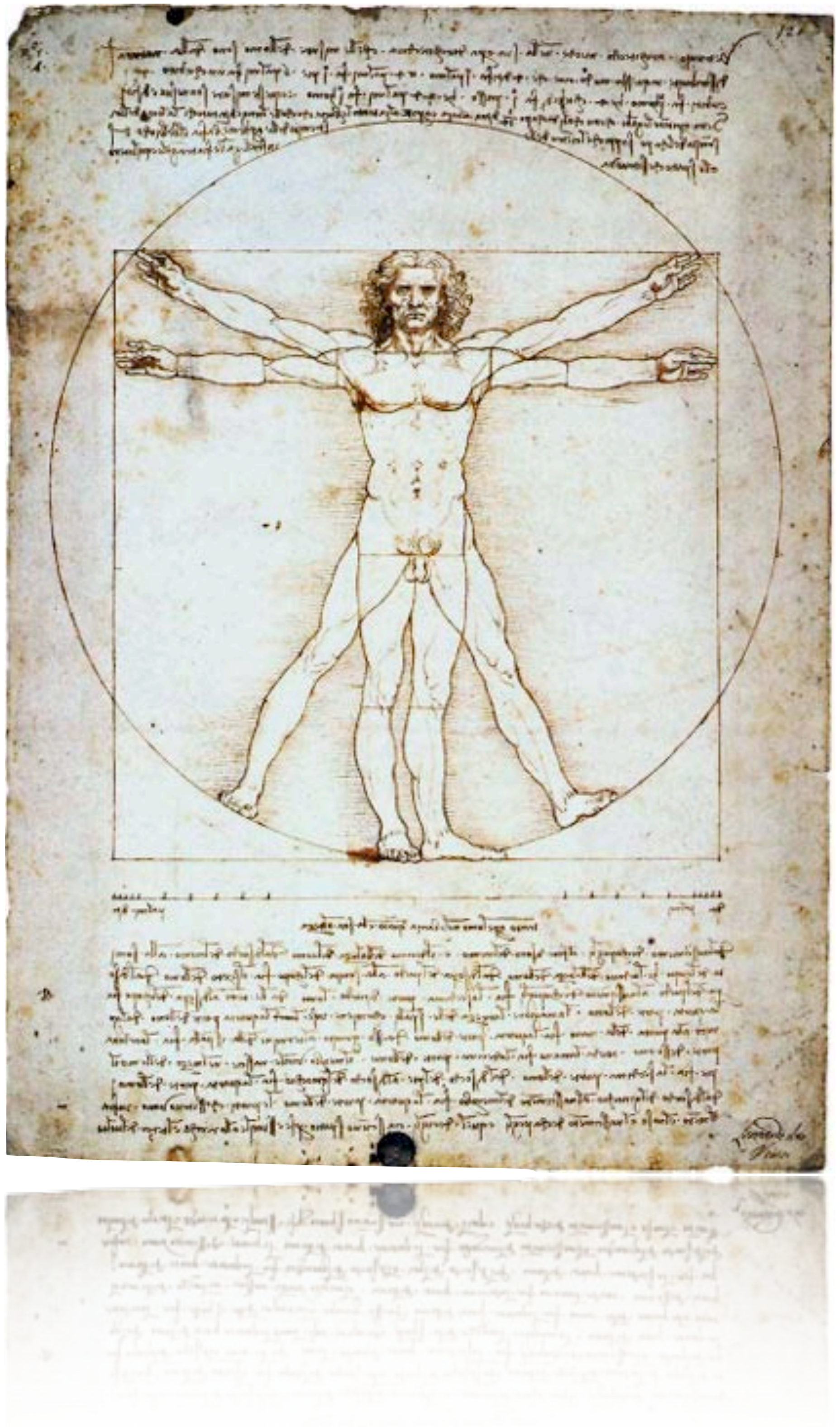
The Scent of Symmetry: A Human Sex Pheromone that Signals Fitness?

Randy Thornhill
Department of Biology, The University of New Mexico, Albuquerque, New Mexico
Steven W. Gangestad
Department of Psychology, The University of New Mexico, Albuquerque, New Mexico

A previous study by the authors showed that the body scent of men who have greater body bilateral symmetry is rated as more attractive by normally ovulating (non-pill-using) women during the period of highest fertility based on day within the menstrual cycle. Women in low-fertility phases of the cycle and women using hormone-based contraceptives do not show this pattern. The current study replicated these findings with a larger sample and statistically controlled for men's hygiene and other factors that were not controlled in the first study. The current study also examined women's scent attractiveness to men and found no evidence that men prefer the scent of symmetric women. We propose that the scent of symmetry is an honest signal of phenotypic and genetic quality in the human male, and chemical candidates are discussed. In both sexes, facial attractiveness (as judged from photos) appears to predict body scent attractiveness to the opposite sex. Women's preference for the scent associated with men's facial attractiveness is greatest when their fertility is highest across the menstrual cycle. The results overall suggest that women have an evolved preference for sires with good genes. **© 1999 Elsevier Science Inc.**

KEY WORDS: Androgens; Androstenone; Androstenol; Developmental instability; Fluctuating asymmetry; Handicap theory; Mate choice; Menstrual cycle; Pheromones; Sexual selection; Signaling.

Evolution and Human Behavior 20: 175–201 (1999)
© 1999 Elsevier Science Inc. All rights reserved.
655 Avenue of the Americas, New York, NY 10010



Fluctuating Asymmetry (FA)

Composite = FA

“Results indicated that normally cycling (non-pill using) women near the peak fertility of their cycle tended to prefer the scent of shirts worn by symmetrical men.”

“Women with partners possessing low FA and their partners reported significantly more copulatory female orgasms than were reported by women with partners possessing high FA and their partners.”

“...there is a real, common, causal link between bodily asymmetry and lowered IQ.”

“Breast asymmetry is likely to be a predictor of, rather than the effect of breast cancer.”

“Subjects who had few or no sperm in their ejaculates tended to have high FA.”

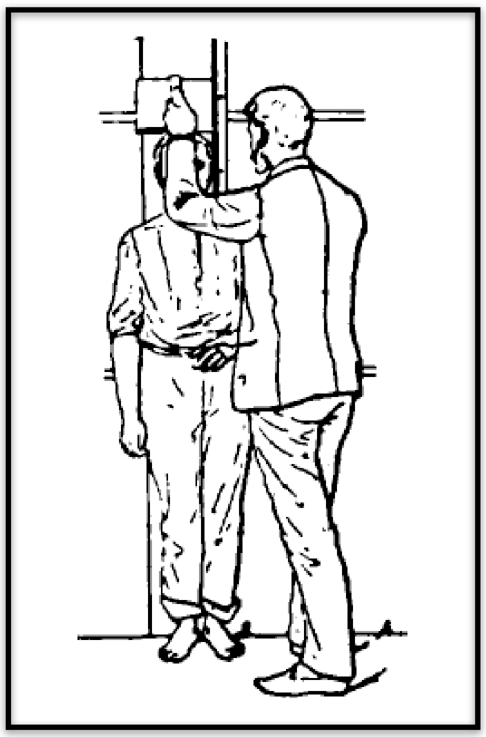
“We found the [Beck Depression Index, BDI] was positively related to fluctuating asymmetry in men but not women”

“We conclude that symmetry in traits such as nostrils and ears indicates good running ability. It may therefore be useful in predicting the future potential of young athletes.”

$$C = \sum w_i Y_i = w_1 Y_1 + w_2 Y_2 + \cdots + w_p Y_p$$

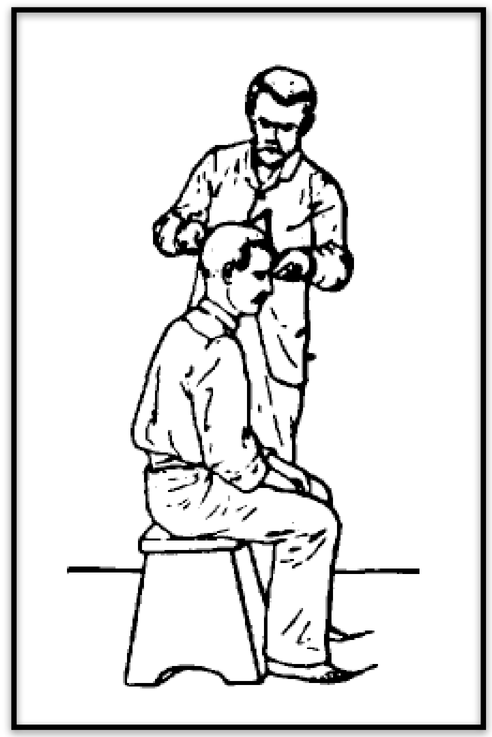
Y_1, Y_2, etc. are scores on existing variables

w_1, w_2, etc. are weights for each variable



$Y_1 = 178.4$

$w_1 = 1$



$Y_2 = 28.1$

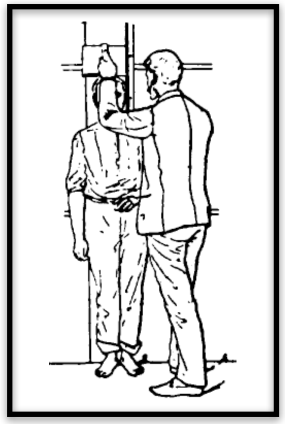
$w_2 = 2$

$Y_3 = 7.3$

$w_3 = -1$

$$C = (1 \times 178.4) + (2 \times 28.1) + (-1 \times 7.3)$$

$C = 227.3 \longleftarrow$ Supervariable = Stature

Person	Aspect 1 Y_1	Aspect 2 Y_2	Aspect 3 Y_3	C_1 $(1, 2, -1)$	C_2 $(2, -2, 1)$
1	178.4	28.1	7.3	227.3	307.9
2	167.0	24.7	6.7	209.8	291.3
3	170.2	27.6	5.8	219.6	291.1
4	187.9	29.3	8.5	237.8	325.8
5	175.2	30.9	6.2	230.9	294.8

Linear composites are used to convert multivariate relationships into bivariate relationships.

For example, start with a multivariate relationship:

$$Y \leftarrow X_1, X_2, X_3$$

Create a linear composite:

$$X_1, X_2, X_3 \Rightarrow C_1$$

resulting in a bivariate relationship:

$$Y \leftarrow C_1$$

Properties of linear composites

- In data analysis, linear composites need to have *specific properties*.
- Weights are calculated mathematically to produce a linear composite with the right properties.
- Different multivariate methods use linear composites with different properties.

Linear Composites in Discriminant Analysis

Discriminant Analysis looks for a relationship between a **categorical** variable and a set of variables:

$$X_{cat} \leftarrow Y_1, Y_2, Y_3$$

Pick some weights: w_1, w_2, w_3

Create a linear composite:

$$C_1 = w_1 Y_1 + w_2 Y_2 + w_3 Y_3$$

Resulting in a **t-test** or **F-test**:

$$X_{cat} \leftarrow C_1$$

Bald

Hairy

Hair Density (2 Levels)	$\leftarrow$	Entry GPA	Gene Quality	Hand Span
X_{cat}	$\leftarrow$	Y_1	Y_2	Y_3

	GPA	Gene Quality	Hand Span
	6	2	10
	4	3	9
	4	4	11
	7	2	11
	5	2	10
Mean	5.2	2.6	10.2
	7	3	8
	5	3	7
	4	4	9
	8	2	8
	5	2	5
Mean	5.8	2.8	7.4
t-value	-0.64	-0.37	3.61

By doing three t-tests (one for each variable), the three variables may be correlated.

So the interpretations of the t-tests are not independent (i.e., we aren't properly assessing the effect of GPA and gene quality *independently* because GPA and gene quality may, in fact, be correlated.)

Another approach is to combine the three variables into a composite variable and perform a t-test on this composite variable.

But how do we combine the scores?

	w_1	w_2	w_3
C_1	1	1	1
C_2	1	2	-1
C_3	1	1	-2

*These weights are arbitrary in this example. Later, we'll cover how to find optimal weights.

	GPA	Gene Quality	Hand Span	C_1 $(1,1,1)$	C_2 $(1,2,-1)$	C_3 $(1,1,-2)$
	6	2	10	18	0	-12
	4	3	9	16	1	-11
	4	4	11	19	1	-14
	7	2	11	20	0	-13
	5	2	10	17	-1	-13
Mean	5.2	[illegible]	10.2	18	0.2	-12.6
	[illegible]	[illegible]	[illegible]	18	5	-6
	[illegible]	[illegible]	[illegible]	15	4	-6
	[illegible]	[illegible]	[illegible]	17	3	-10
	[illegible]	[illegible]	[illegible]	18	4	-6
	[illegible]	[illegible]	[illegible]	12	4	-3
Mean	5.[illegible]	[illegible]	[illegible]	16	4	-6.2
t-value	-0.64	-0.37	3.61	1.49	-7.76	5.23

The goal here is to find the linear composite such that the t-value for the differences between the groups is as large as possible. The weights give the 'relative importance' of the variables. The optimum weights depend essentially on the pattern of correlations among the variables.

Bald

With more than two groups, a t-value is no longer appropriate. Instead, an F-value is the appropriate index of between-group differences. The goal now would be to find the linear composite such that the F-value for the difference between groups is as large as possible.

airy

Not Quite Bald

Linear Composites in Multiple Regression

Multiple Regression looks for a relationship between a **continuous** variable and a set of variables:

$$Y_{cont} \leftarrow X_1, X_2, X_3$$

Pick some weights: a_1, a_2, a_3

Create a linear composite:

$$C_1 = a_1 X_1 + a_2 X_2 + a_3 X_3$$

Resulting in a **correlation**:

$$Y_{cont} \leftarrow C_1$$

Final Honours Grade	$\leftarrow$	Entry GPA	Time to Run 5 kilometres	Daily Caffeine Intake
Y	$\leftarrow$	X_1	X_2	X_3

	Final Honours Grade	Entry GPA	Time to Run 5 km	Daily Caffeine Intake
	87.3	6.9	19.0	476.5
	72.4	6.4	43.4	663.3
	56.1	5.2	28.6	383.9
	66.2	6.1	41.0	546.8
	53.0	6.0	45.7	422.0
	61.9	6.9	38.8	473.0
Mean	66.2	6.3	36.1	494.3
r		0.39	-0.59	0.46

By computing three correlation coefficients for each predictor and the criterion, the three predictors may be correlated.

So the interpretations of the simple correlations are not independent.

Another approach is to combine the three predictors into a composite and perform a correlation with the composite variable and the criterion.

But how do we combine the scores?

	a_1	a_2	a_3
C_1	4	2	3
C_2	-4	2	3
C_3	1	1	1

*Again, these weights are arbitrary. Notice that we're using different symbols to indicate that across different multivariate methods, different notation systems are traditionally used.

	Final Honours Grade	Entry GPA	Time to Run 5 km	Daily Caffeine Intake	C_1 $(4, 2, 3)$	C_2 $(-4, 2, 3)$	C_3 $(1, 1, 1)$
	87.3	6.9	19.0	476.5	1495.1	1439.7	502.4
	72.4	6.4	43.4	663.3	2102.5	2051.1	713.2
	56.1	5.2	28.6	383.9	1229.4	1188.1	417.6
	66.2	6.1	41.0	546.8	1746.9	1697.8	593.9
	53.0	6.0	45.7	422.0	1381.4	1333.4	473.7
	61.9	6.9	38.8	473.0	1524.3	1469.0	518.7
Mean	66.2	6.3	36.1	494.3	1579.9	1529.9	536.6
r		0.39	0.59	0.46	0.41	0.4	0.38

The goal here is to find the linear composite such that the correlation with the criterion is as large as possible. The weights give the ‘relative importance’ of the variables. The optimum weights depend essentially on the pattern of correlations among the variables.

Linear Composites in Factor Analysis

Factor Analysis looks for a single variable that summaries multiple variables, without losing too much information (variance).

$$V_1, V_2, V_3$$

Pick some weights: a_1, a_2, a_3

Create a linear composite:

$$C_1 = a_1 V_1 + a_2 V_2 + a_3 V_3$$

Resulting in a **single summary variable**: C_1

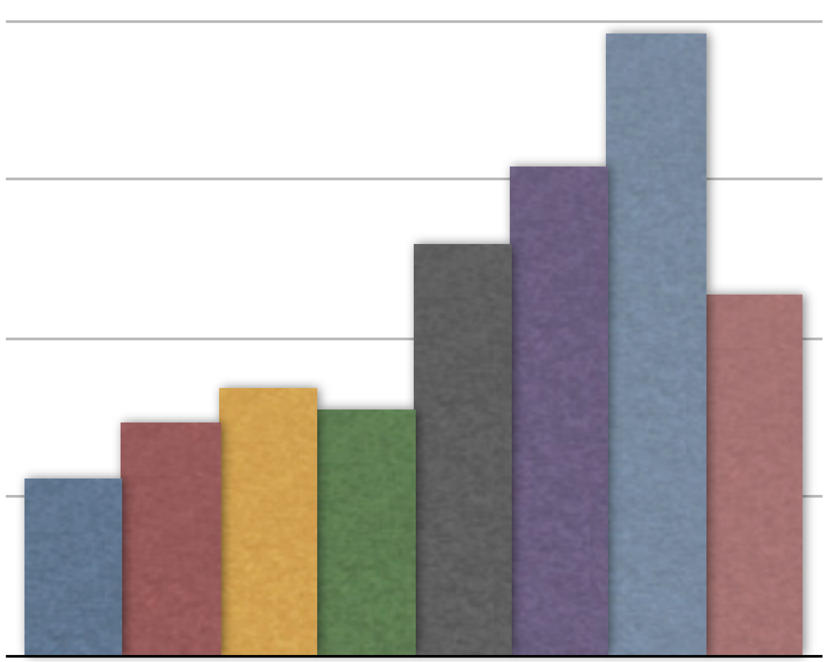
C_1 has a Variance that captures the ‘information’.

Measures that ‘define’ success...

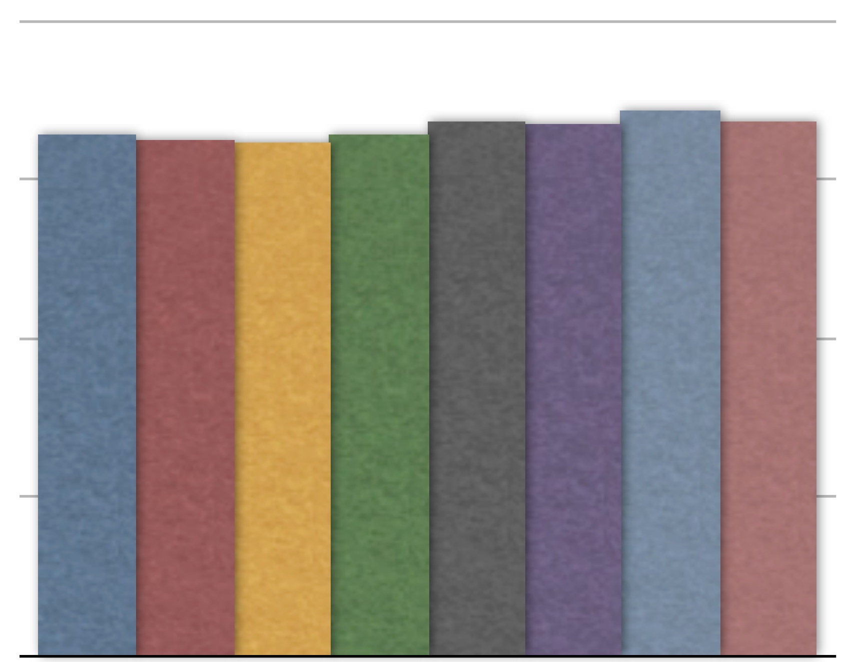
Typing Speed	Emotional Stability	Chess Experience
V_1	V_2	V_3

...but how do we know whether we have a ‘good’ measure?

One criterion for a ‘good’ variable is that is serves to distinguish between cases.



Good



Not so good

	Typing Speed	Emotional Stability	Chess Experience
	2	4	5
	1	7	2
	9	0	5
	6	2	4
	2	6	3
Variance	11.5	8.2	1.7

By computing the variance for each measure, the three measures may be correlated.

So the interpretations of the measures are not independent.

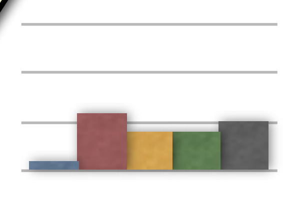
Another approach is to combine the three measures into a composite and compute the variance of the composite variable.

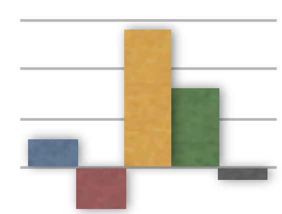
But how do we combine the scores?

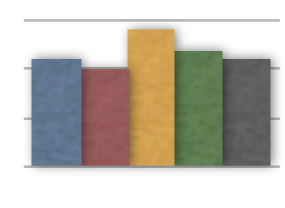
	a_1	a_2	a_3
C_1	1	1	-1
C_2	1	-1	1
C_3	1	1	1

*Note that the variance of the linear composite can get large if we change the magnitude of the weights. So the weights are constrained so that their sums of squares are equal.

	Typing Speed	Emotional Stability	Chess Experience	C_1 $(1, 1, -1)$	C_2 $(1, -1, 1)$	C_3 $(1, 1, 1)$
	2	4	5	1	3	11
	1	7	2	6	-4	10
	9	0	5	4	14	14
	6	2	4	4	8	12
	2	6	3	5	-1	11
Variance	11.5	8.2	1.7	3.5	51.5	2.3







The goal here is to find the linear composite such that the scatter (spread) of the scores is a large as possible. That is, the linear composite has the largest possible variance. This gives the ‘most important factor’. The optimum weights depend essentially on the pattern of correlations among the variables.

The ‘trick’ used to handle multiple variables is to ‘add’ them up to form one variable (the linear composite) and then to perform the familiar univariate analyses.

In data analysis, linear composites are created with specific properties in order to maximise something:

- in *discriminant analysis*, create linear composites to maximise **group differences** (or a t-value or an F-value).
- in *multiple regression*, create linear composites to maximise a **correlation**.
- in *factor analysis*, create linear composites to maximise a **variance.**

Questions

1. Explain how linear composites are used in multiple regression. Could simple correlations between the criterion and the predictors give the same information as using a linear composite?

2. Why is variance an important concept?

3. Explain why separate univariate (bivariate) analyses are not appropriate for handling multivariate data.

4. Why is the linear composite formed to maximise something? Why not just have arbitrary combinations of the measured variables?

5. Using the body measurement example above (e.g., 178.4, 28.1,..., 6.2) apply the weights used in the factor analysis example to see which of the three linear composites best maximised the variance.

Multiple Regression: An Overview

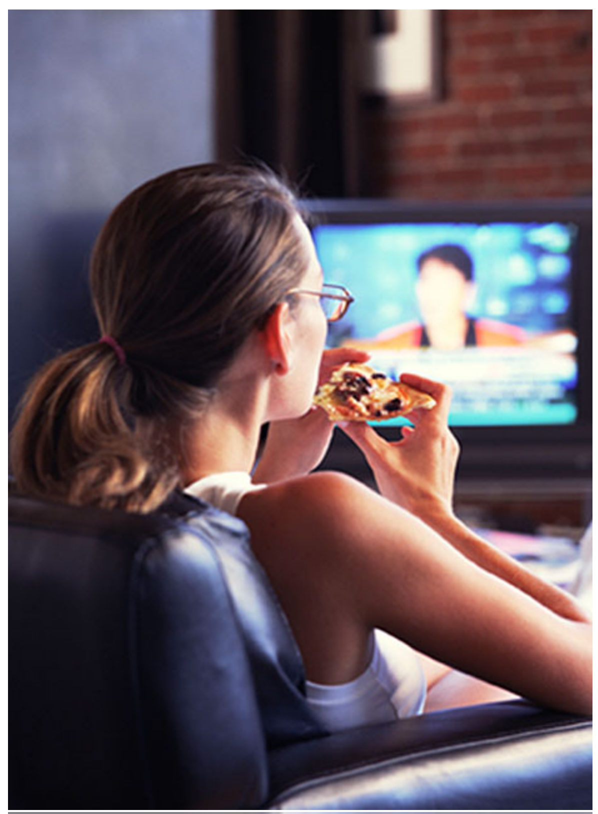
Major themes in multiple regression:

- Data = Model + Residual
  - The model is specified by the weights for the linear combination of variables.
- Sums of squares and variances can be partitioned.
- Estimating the model for the data.
- How well does the model fit the data?
  - Statistical testing
- Can we trust the model?

Motivational Example

Segrin & Nabi (2002)

What's the relationship between watching television, holding idealistic expectations about marriage, and intentions to marry?



N=285 never-married University students.

Person	overall tv viewing	romantic tv viewing	idealistic marriage expectation	intention to marry	Age	Gender
1	2.56	4.75	3.95	4.5	18	F
2	4.61	2.34	2.87	3.01	22	M
⋮	⋮	⋮	⋮	⋮	⋮	⋮
285	1.41	1.05	1.53	1.25	44	M

Intentions to Marry	$\leftarrow$	Television Viewing	Holding idealistic expectations about marriage
Y	$\leftarrow$	X_1	X_2

Results from regression and path analyses indicate that, although overall television viewing has a negative association with idealistic marriage expectations, viewing of romantic genre programming (e.g., romantic comedies, soap operas) was positively associated with idealistic expectations about marriage. Further, a strong and positive association between these expectations and marital intentions was evidenced.

Bivariate

Simple Linear Regression

$$Y \leftarrow X$$

Independent Groups
t-test

$$Y \leftarrow X$$

Two variable Correlation

$$Y \leftrightarrow X$$

Multivariate

Two Predictor
Multiple Regression

$$Y \leftarrow X_1, X_2$$

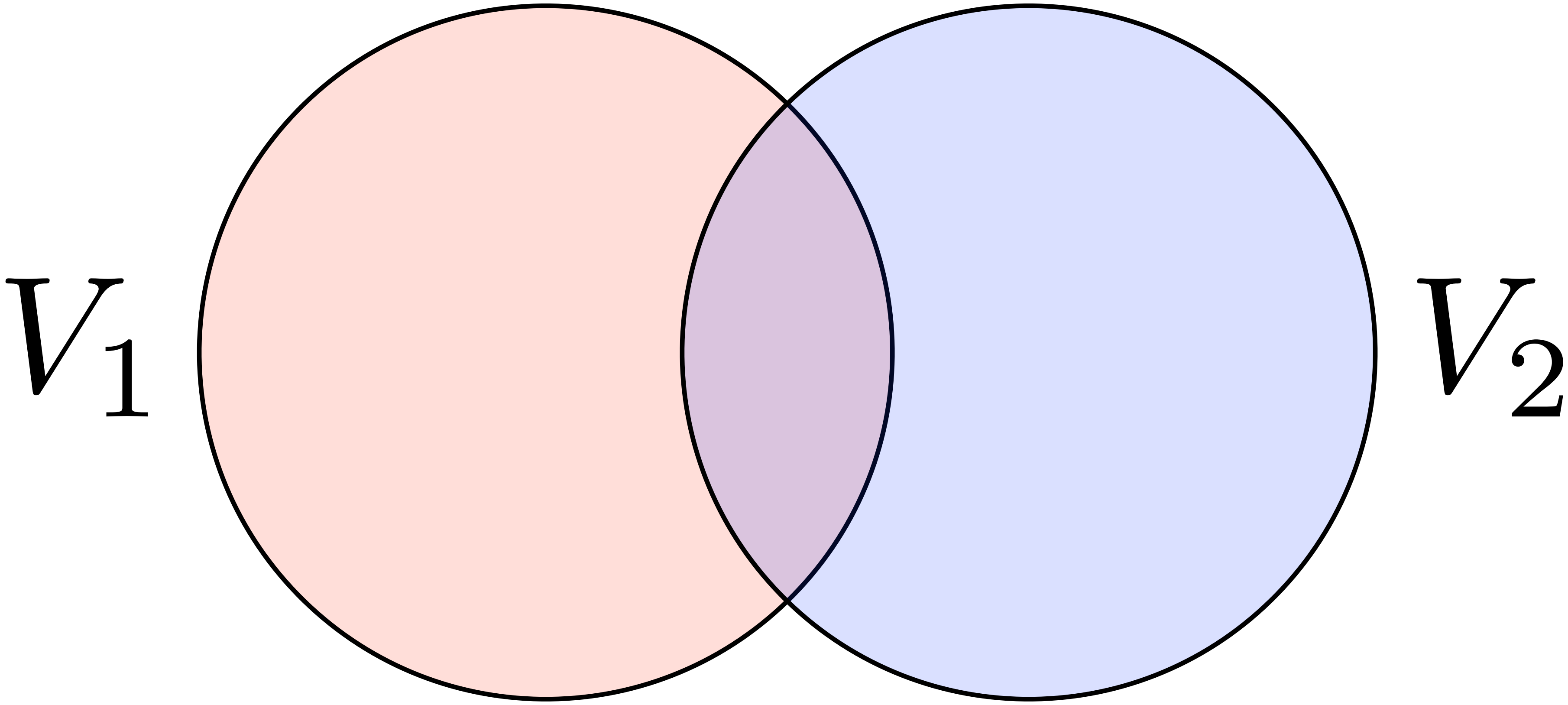
Many Predictor
Multiple Regression

$$Y \leftarrow X_1, X_2 \ldots X_q$$

Correlation

A correlation (r) represents the degree of a linear relationship between two variables, ranging from –1 to +1.

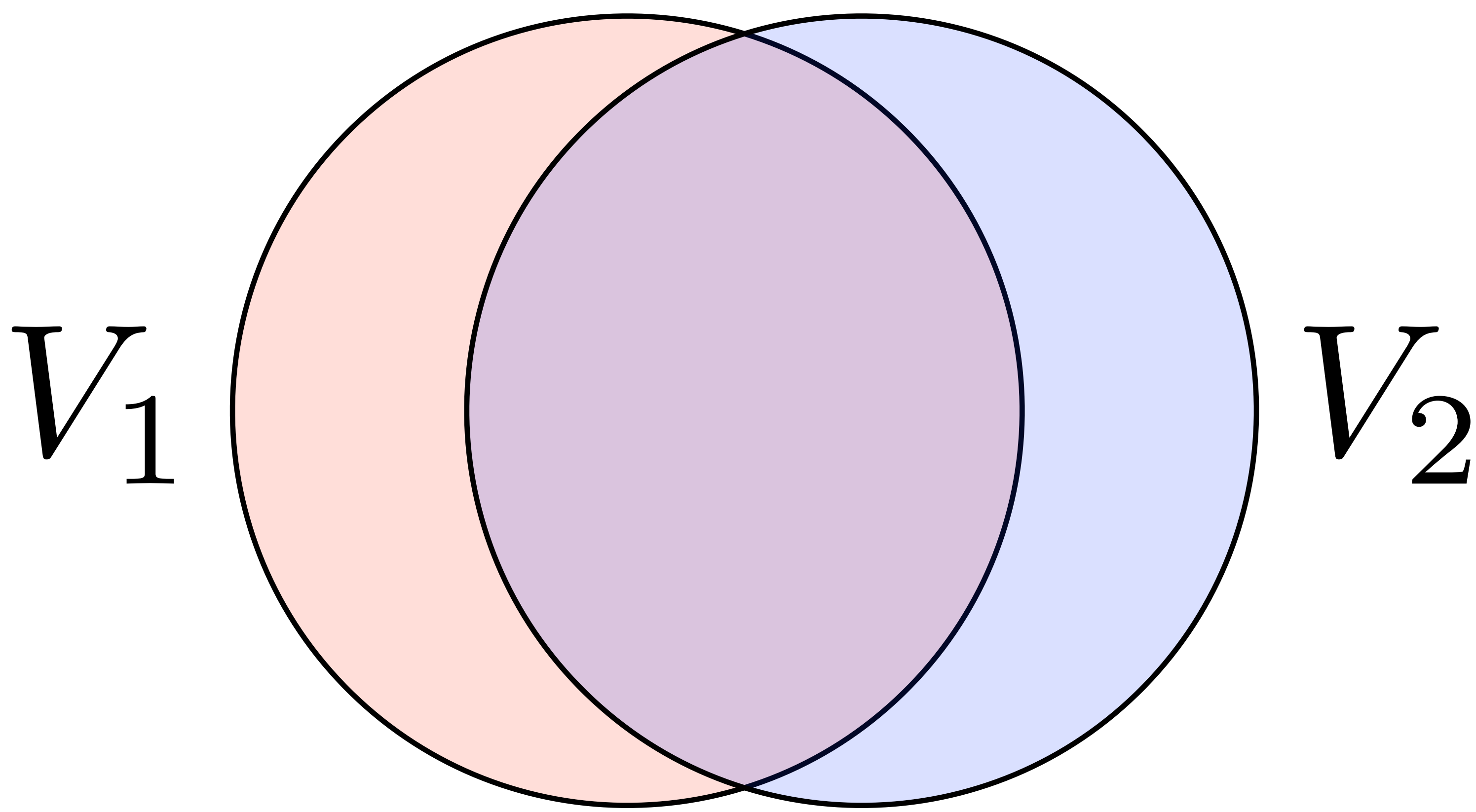
It forms the basis for all other multivariate methods in this course.



a low correlation

$r = .3$

9% overlapping variance

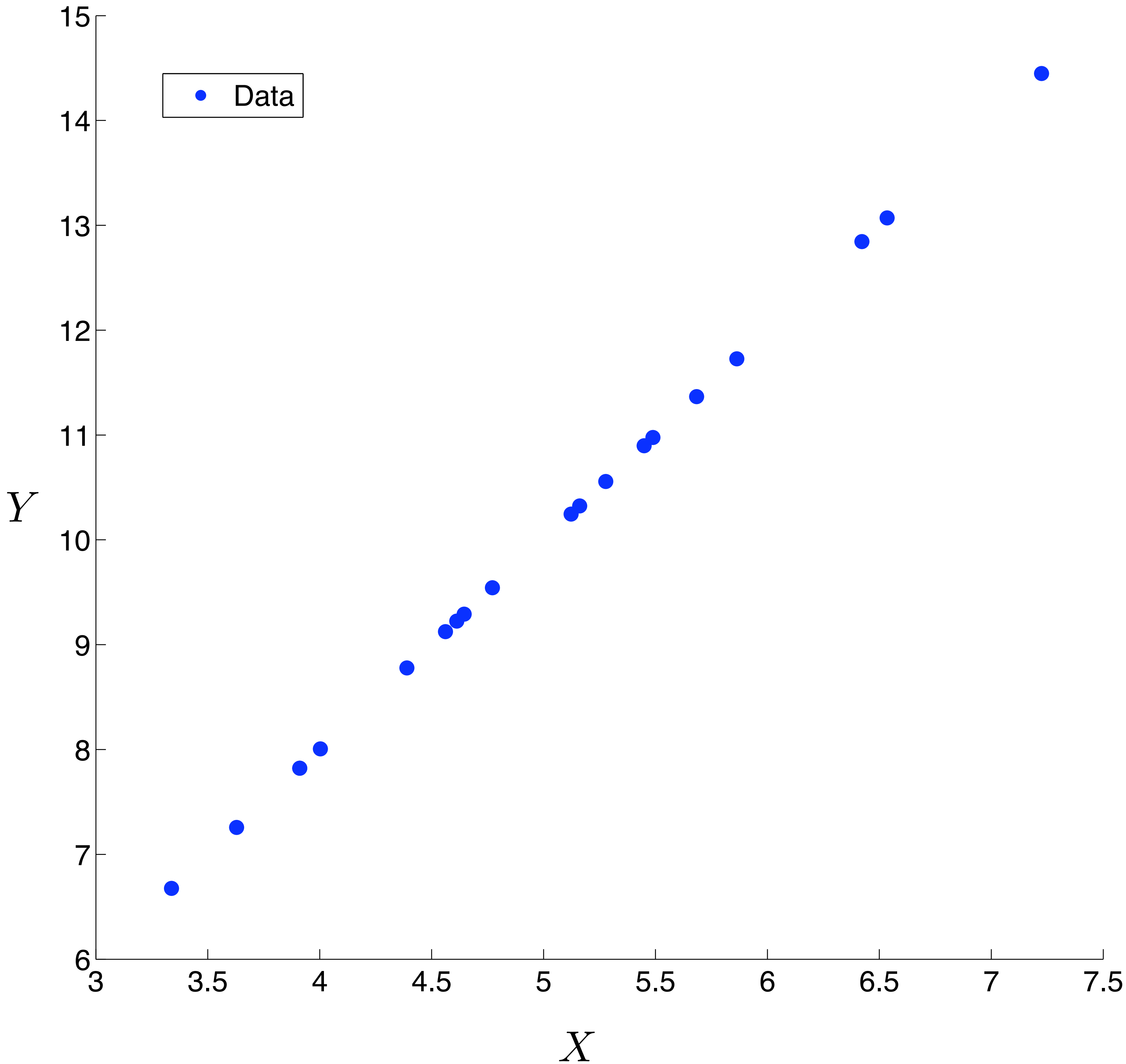


a moderate correlation

$r = .8$

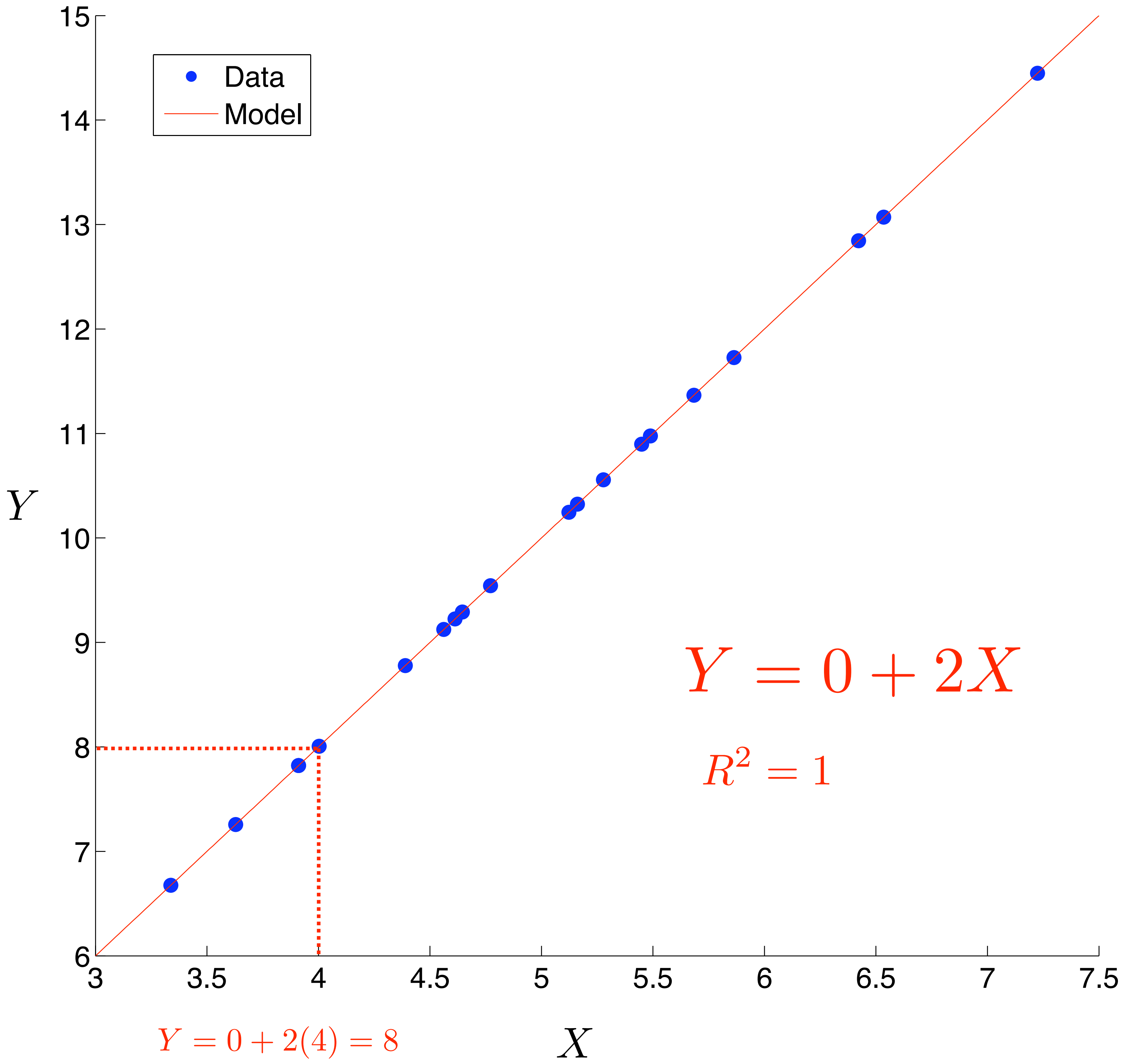
64% overlapping variance

$$Y \leftarrow X$$



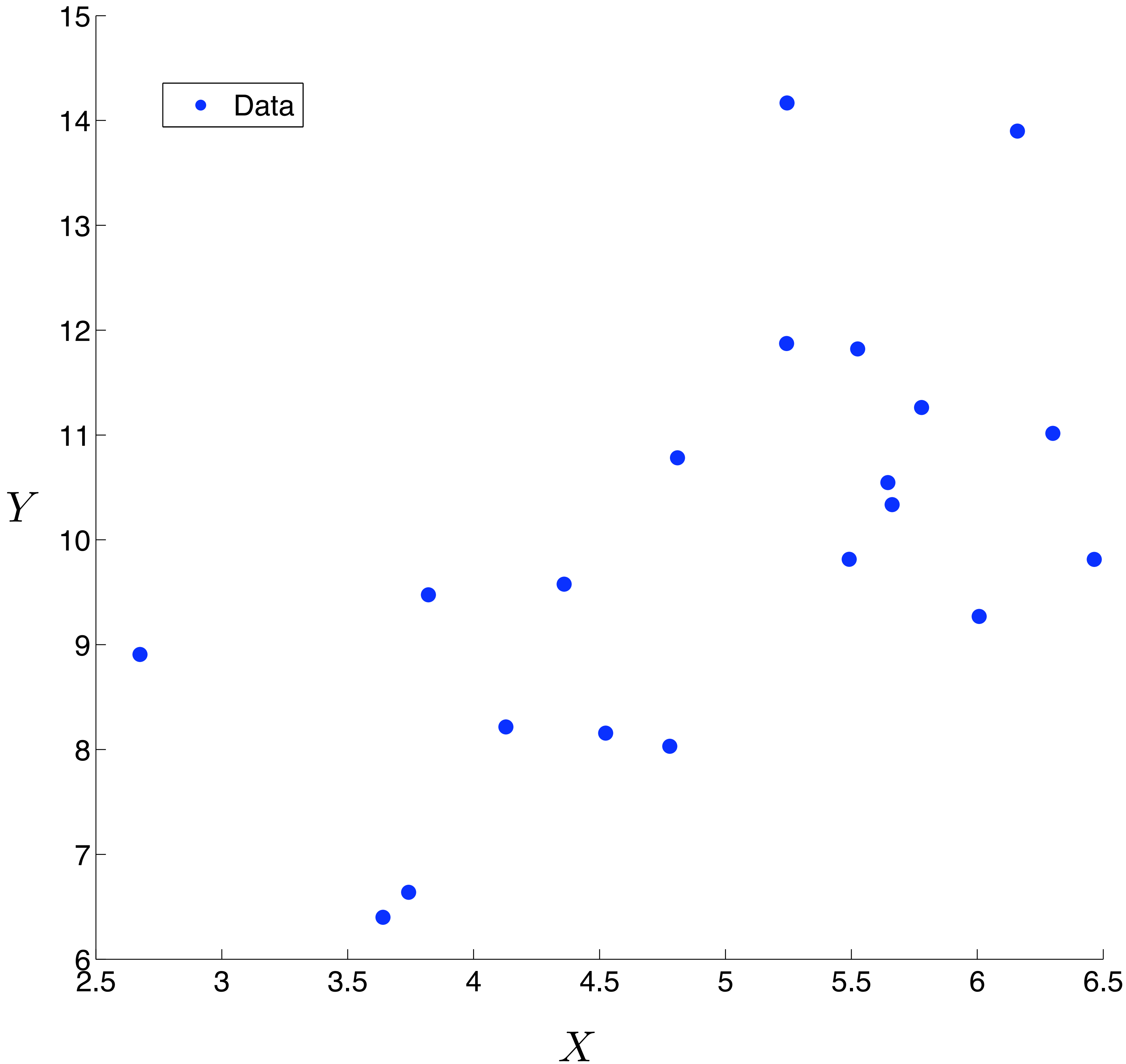
X	Y
4.61	9.22
4.65	9.29
5.12	10.25
5.28	10.56
5.86	11.73
6.42	12.84
4.56	9.12
6.54	13.07
4.39	8.78
3.63	7.26
5.16	10.32
5.68	11.37
4.77	9.54
5.49	10.98
3.91	7.82
5.45	10.90
3.34	6.68
4.00	8.01
3.91	7.82
7.22	14.45

$$Y \leftarrow X$$



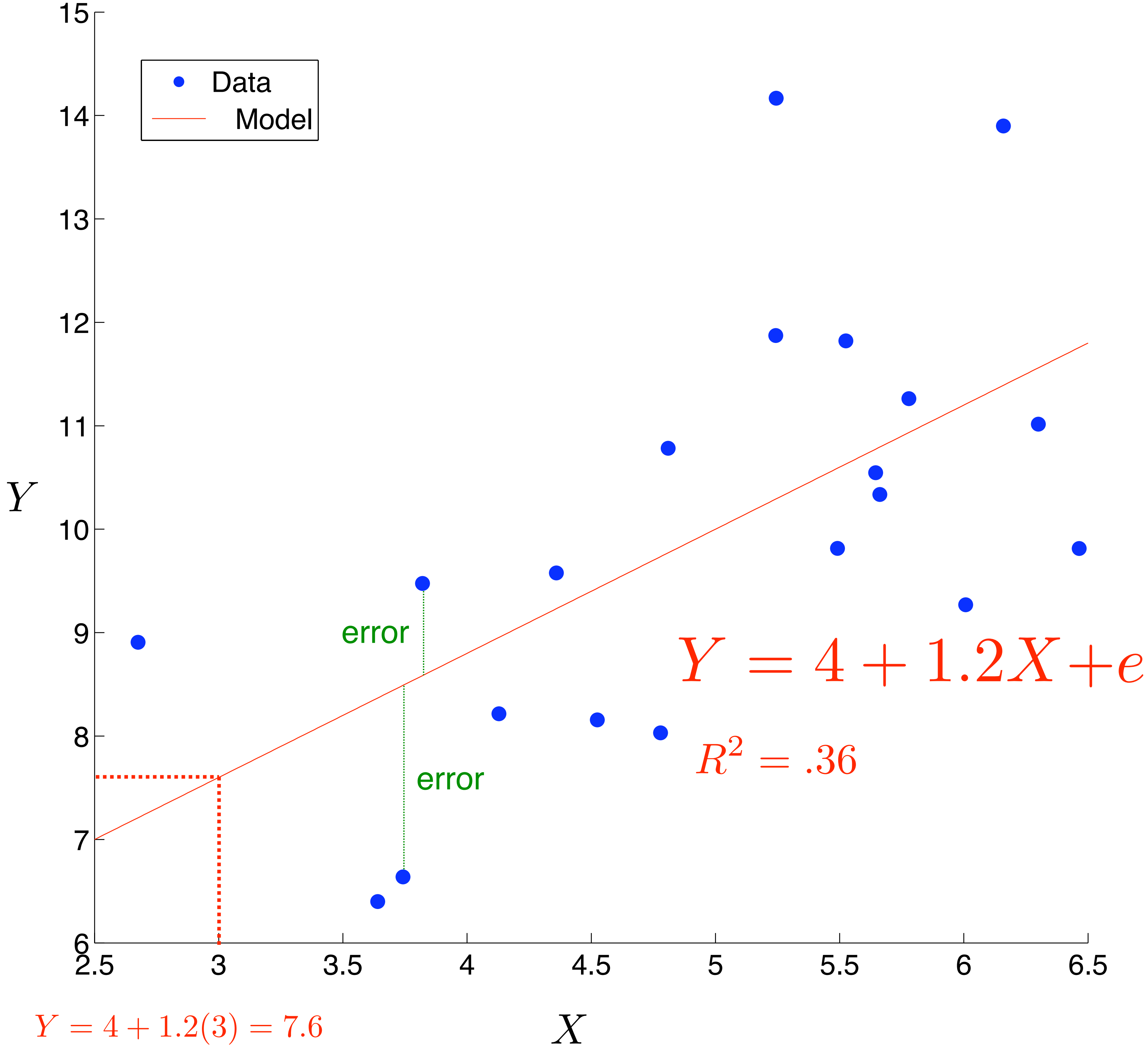
X	Y
4.61	9.22
4.65	9.29
5.12	10.25
5.28	10.56
5.86	11.73
6.42	12.84
4.56	9.12
6.54	13.07
4.39	8.78
3.63	7.26
5.16	10.32
5.68	11.37
4.77	9.54
5.49	10.98
3.91	7.82
5.45	10.90
3.34	6.68
4.00	8.01
3.91	7.82
7.22	14.45

$$Y \leftarrow X$$



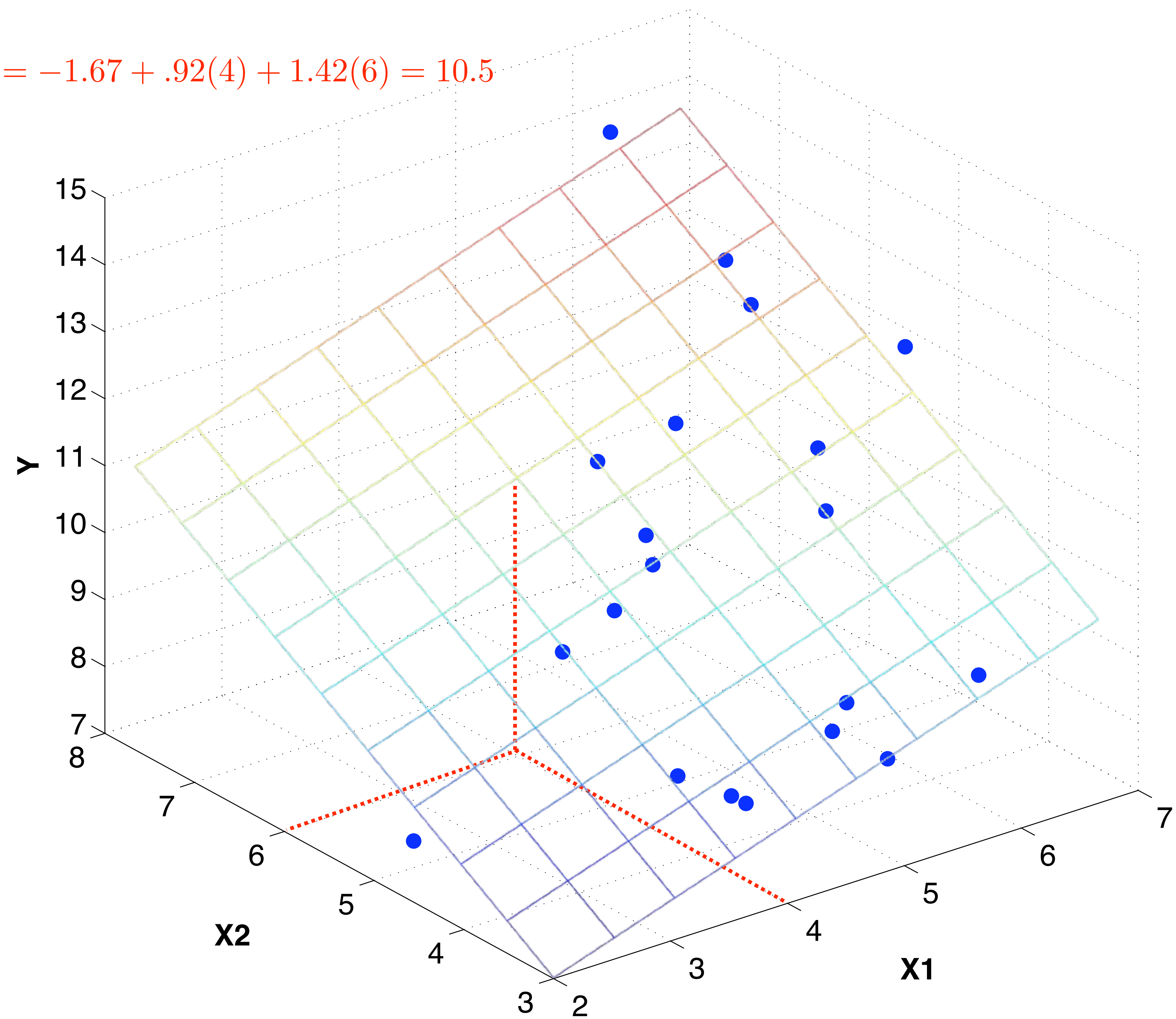
X	Y
5.53	11.82
6.46	9.81
5.24	11.87
4.78	8.03
4.52	8.16
2.68	8.91
4.13	8.22
4.36	9.58
6.30	11.02
5.78	11.26
5.66	10.34
3.82	9.48
5.65	10.55
6.01	9.27
5.49	9.81
6.16	13.90
5.24	14.17
3.74	6.64
4.81	10.78
3.64	6.40

$$Y \leftarrow X$$



X	Y
5.53	11.82
6.46	9.81
5.24	11.87
4.78	8.03
4.52	8.16
2.68	8.91
4.13	8.22
4.36	9.58
6.30	11.02
5.78	11.26
5.66	10.34
3.82	9.48
5.65	10.55
6.01	9.27
5.49	9.81
6.16	13.90
5.24	14.17
3.74	6.64
4.81	10.78
3.64	6.40

$$Y \leftarrow X_1, X_2$$



X_1	X_2	Y
4.10	5.06	9.81
5.34	4.09	8.44
5.45	4.46	10.97
4.39	3.97	7.56
4.30	4.01	7.69
5.92	3.37	9.06
4.87	5.71	10.02
3.86	5.33	9.13
5.03	5.59	11.69
5.06	5.89	9.34
5.51	3.85	7.69
5.91	5.89	12.75
2.18	4.80	7.63
4.47	5.73	11.33
6.19	6.54	12.78
6.81	5.35	12.01
4.16	4.43	7.76
4.93	3.71	8.53
5.73	4.92	11.42
5.78	7.29	14.37

$$Y = -1.67 + .92(X_1) + 1.42(X_2)$$

$$R^2 = .84$$

The notion of modelling the data

$$Y \leftarrow X$$

$$Y_i = a + bX_i + e_i$$

Actual Score = Predicted from Model + Error

$$Y_i = Y_i' + e_i$$

Actual Score = Predicted Score + Residual

Actual Score = Regression + Residual

In general:

DATA = MODEL + RESIDUAL

The notion of modelling the data

$$Y \leftarrow X_1, X_2$$

$$Y_i = a + b_1 X_{1i} + b_2 X_{2i} + e_i$$

Actual Score = Predicted from Model + Error

$$Y_i = Y_i' + e_i$$

Actual Score = Predicted Score + Residual

Actual Score = Regression + Residual

In general:

DATA = MODEL + RESIDUAL

The general situation

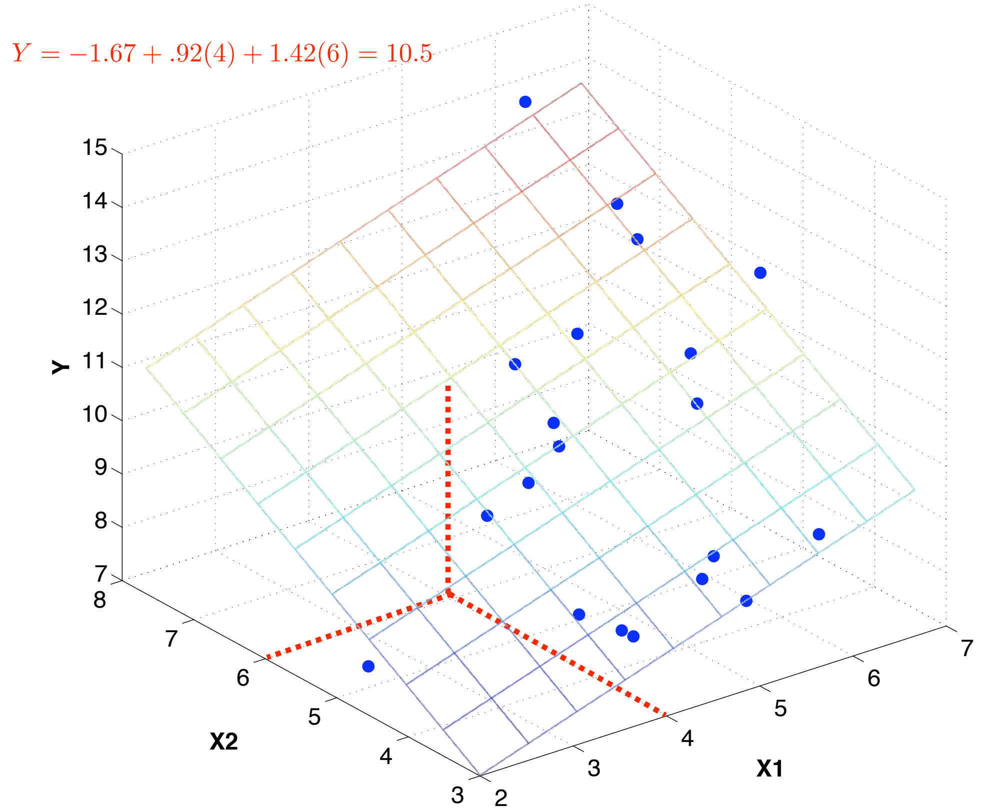
$$Y = \boxed{X_1, X_2 \ldots X_p} + e$$

$$Y = \boxed{Y'} + e$$

Y' is a linear composite and represents the model of the data

DATA = MODEL + RESIDUAL

$$Y \leftarrow X_1, X_2$$



$$Y = -1.67 + .92(X_1) + 1.42(X_2)$$
$R^2 = .84$

X_1	X_2	Y
4.10	5.06	9.81
5.34	4.09	8.44
5.45	4.46	10.97
4.39	3.97	7.56
4.30	4.01	7.69
5.92	3.37	9.06
4.87	5.71	10.02
3.86	5.33	9.13
5.03	5.59	11.69
5.06	5.89	9.34
5.51	3.85	7.69
5.91	5.89	12.75
2.18	4.80	7.63
4.47	5.73	11.33
6.19	6.54	12.78
6.81	5.35	12.01
4.16	4.43	7.76
4.93	3.71	8.53
5.73	4.92	11.42
5.78	7.29	14.37

These best fitting regression coefficients produce a prediction equation for which squared differences between Y and Y' are at a minimum.

Because the squared error of prediction $(Y - Y')^2$ are minimised, this solution is called a least-squares solution.

	X_1	X_2	Y	Y'	$Y-Y'$	$(Y-Y')^2$	$Y-\overline{Y}$	$(Y-\overline{Y})^2$	$Y'-\overline{Y}$	$(Y'-\overline{Y})^2$
	4.10	5.06	9.81	9.26	0.54	0.30	-0.19	0.04	-0.74	0.55
	5.34	4.09	8.44	9.02	-0.58	0.33	-1.56	2.42	-0.98	0.96
	5.45	4.46	10.97	9.65	1.32	1.74	0.97	0.95	-0.35	0.12
	4.39	3.97	7.56	7.99	-0.43	0.19	-2.44	5.97	-2.01	4.04
	4.30	4.01	7.69	7.96	-0.26	0.07	-2.31	5.31	-2.04	4.18
	5.92	3.37	9.06	8.54	0.52	0.27	-0.94	0.89	-1.46	2.13
	4.87	5.71	10.02	10.88	-0.86	0.74	0.02	0.00	0.88	0.78
	3.86	5.33	9.13	9.42	-0.29	0.08	-0.87	0.76	-0.58	0.34
	5.03	5.59	11.69	10.86	0.83	0.69	1.69	2.87	0.86	0.74
	5.06	5.89	9.34	11.31	-1.97	3.88	-0.66	0.43	1.31	1.73
	5.51	3.85	7.69	8.83	-1.14	1.29	-2.31	5.32	-1.17	1.37
	5.91	5.89	12.75	12.10	0.65	0.42	2.75	7.55	2.10	4.41
	2.18	4.80	7.63	7.13	0.50	0.25	-2.37	5.61	-2.87	8.21
	4.47	5.73	11.33	10.56	0.77	0.60	1.33	1.77	0.56	0.31
	6.19	6.54	12.78	13.28	-0.50	0.25	2.78	7.73	3.28	10.73
	6.81	5.35	12.01	12.16	-0.15	0.02	2.01	4.03	2.16	4.67
	4.16	4.43	7.76	8.43	-0.66	0.44	-2.24	5.01	-1.57	2.48
	4.93	3.71	8.53	8.11	0.42	0.17	-1.47	2.17	-1.89	3.58
	5.73	4.92	11.42	10.55	0.88	0.77	1.42	2.03	0.55	0.30
	5.78	7.29	14.37	13.96	0.41	0.17	4.37	19.13	3.96	15.71
$\sum$	100.00	100.00	200.00	200.00	0.00	$\boxed{12.67}$	0.00	$\boxed{80.00}$	0.00	$\boxed{67.33}$
MEAN	5.00	5.00	10.00	10.00	0.00	0.63	0.00	4.00	0.00	3.37

$$SS_Y \rightarrow SS_{reg} \quad SS_{res}$$

$$SS_Y = \sum (Y-\overline{Y})^2 \rightarrow SS_{reg} = \sum (Y'-\overline{Y})^2 \quad SS_{res} = \sum (Y-Y')^2$$

$$\boxed{80} \rightarrow \boxed{67.33} \quad \boxed{12.67}$$

$$R^2 = \frac{SS_{reg}}{SS_Y} = \frac{67.33}{80} = .84$$

$$
\begin{array}{cc|c|c|c|}
 & & \multicolumn{2}{c}{\mathbf{R}_{ii}} & \mathbf{R}_{iy} \\
 & & X_1 & X_2 & Y \\
\hline
 & X_1 & 1 & 0.2 & 0.6 \\
\hline
 & X_2 & 0.2 & 1 & 0.8 \\
\hline
\mathbf{R}_{yi} & Y & 0.6 & 0.8 & 1 \\
\hline
\end{array}
$$

$$\mathbf{R}_{ii} = \begin{bmatrix} 1 & .2 \\ .2 & 1 \end{bmatrix} \qquad \mathbf{R}_{ii}^{-1} = \begin{bmatrix} 1.042 & -0.208 \\ -0.208 & 1.042 \end{bmatrix}$$

$$\mathbf{B}_i = \mathbf{R}_{ii}^{-1}\mathbf{R}_{iy}$$

$$\mathbf{B}_i = \begin{bmatrix} 1.042 & -0.208 \\ -0.208 & 1.042 \end{bmatrix} \begin{bmatrix} .6 \\ .8 \end{bmatrix} = \begin{bmatrix} .458 \\ .708 \end{bmatrix}$$

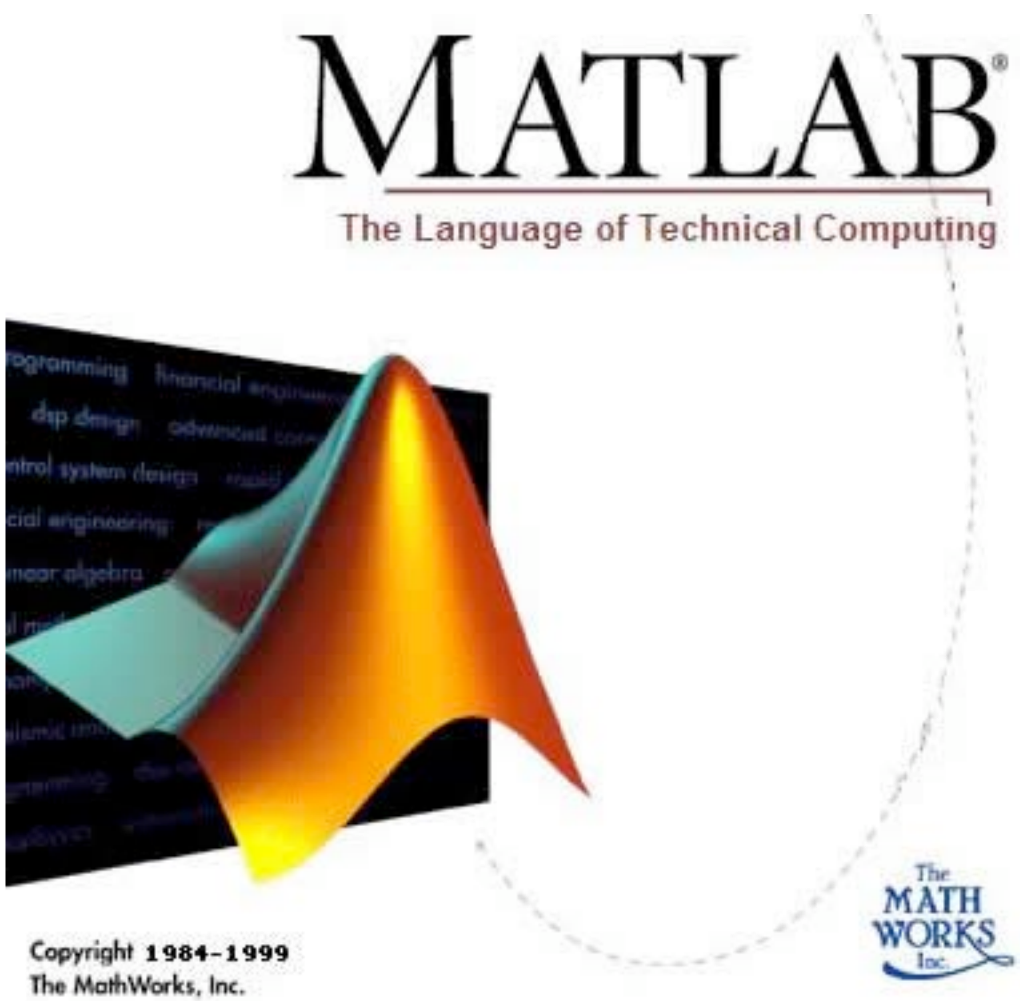
$$\mathbf{R}^2 = \mathbf{R}_{yi}\mathbf{B}_i$$

$$\mathbf{R}^2 = \begin{bmatrix} .6 & .8 \end{bmatrix} \begin{bmatrix} .458 \\ .708 \end{bmatrix}$$

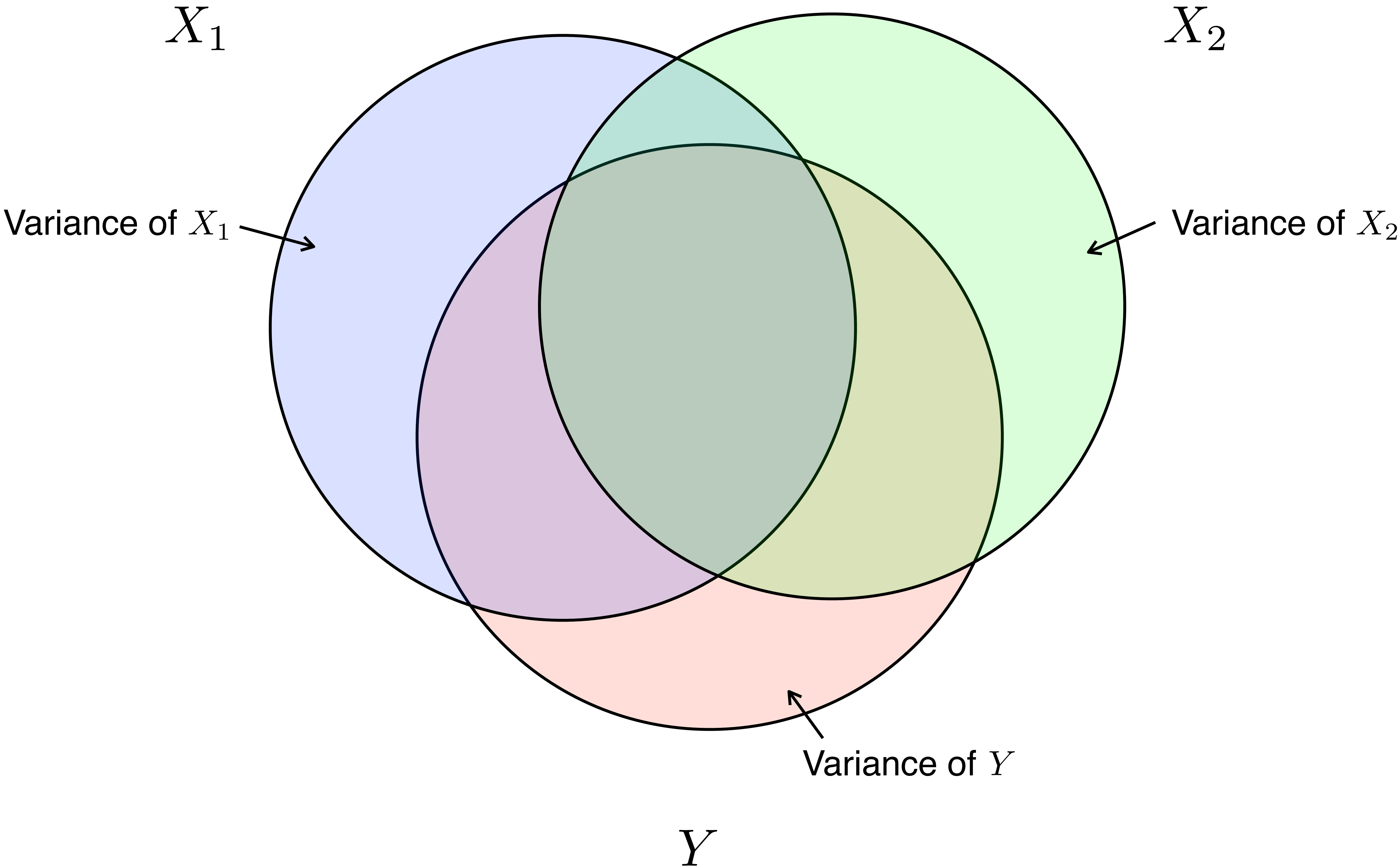
$$\mathbf{R}^2 = 0.84$$

X_1	X_2	Y
4.10	5.06	9.81
5.34	4.09	8.44
5.45	4.46	10.97
4.39	3.97	7.56
4.30	4.01	7.69
5.92	3.37	9.06
4.87	5.71	10.02
3.86	5.33	9.13
5.03	5.59	11.69
5.06	5.89	9.34
5.51	3.85	7.69
5.91	5.89	12.75
2.18	4.80	7.63
4.47	5.73	11.33
6.19	6.54	12.78
6.81	5.35	12.01
4.16	4.43	7.76
4.93	3.71	8.53
5.73	4.92	11.42
5.78	7.29	14.37

For demonstration purposes only!




```
>> R_ii=[1 0.2 ; 0.2 1]

R_ii =

    1.0000    0.2000
    0.2000    1.0000

>> R_iy=[0.6 ; 0.8]

R_iy =

    0.6000
    0.8000

>> R_yi=[0.6 0.8]

R_yi =

    0.6000    0.8000

>> R_ii_inverse=inv(R_ii)

R_ii_inverse =

    1.0417   -0.2083
   -0.2083    1.0417

>> B_i=R_ii_inverse*R_iy

B_i =

    0.4583
    0.7083

>> R2=R_yi*B_i

R2 =

    0.8417
```


X_1
X_2
Variance of X_1
Variance of X_2
Variance of Y
Y

For each case, the score is decomposed into additive components:

$$Y_i = Y_i' + e_i$$

Over cases, variance summarises how much the scores differ from each other:

$$Var(Y) = Var(Y') + Var(e)$$

$$SS_{actual} = SS_{regression} + SS_{residual}$$

Major questions answered by multiple regression

Question 1: Is there an overall relationship between the two predictors and the criterion?

Question 2: Is there a relationship between each *individual* predictor and the criterion? What is the relative importance of each predictor?

Regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	X2, X1[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: Y

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.917[a]	.842	.823	.86319

a. Predictors: (Constant), X2, X1

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	67.333	2	33.667	45.184	.000[a]
	Residual	12.667	17	.745		
	Total	80.000	19			

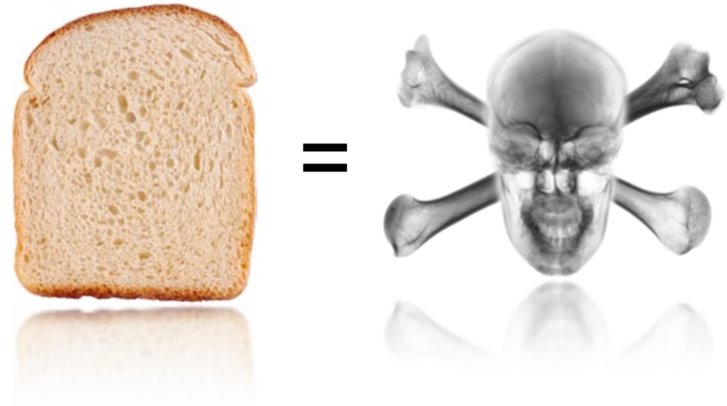
a. Predictors: (Constant), X2, X1

b. Dependent Variable: Y

Coefficients[a]

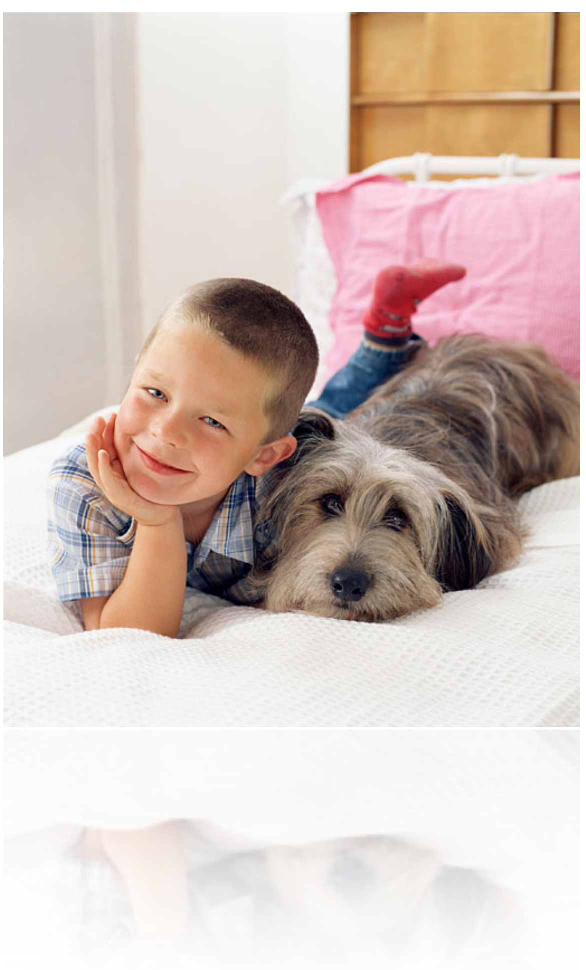
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	−1.667	1.261		−1.322	.204
	X1	.917	.197	.458	4.653	.000
	X2	1.417	.197	.708	7.191	.000

a. Dependent Variable: Y



If you assert from a strong correlation between A and B that A *causes* B, the critic can usually rebut forcefully by proposing some variable C as the underlying causal agent.

... or the cause and effect may be in the reverse direction.



Children with pet dogs are more well behaved than children without pet dogs.

One might conclude that the responsibility of caring for an animal has a maturing influence on the child.

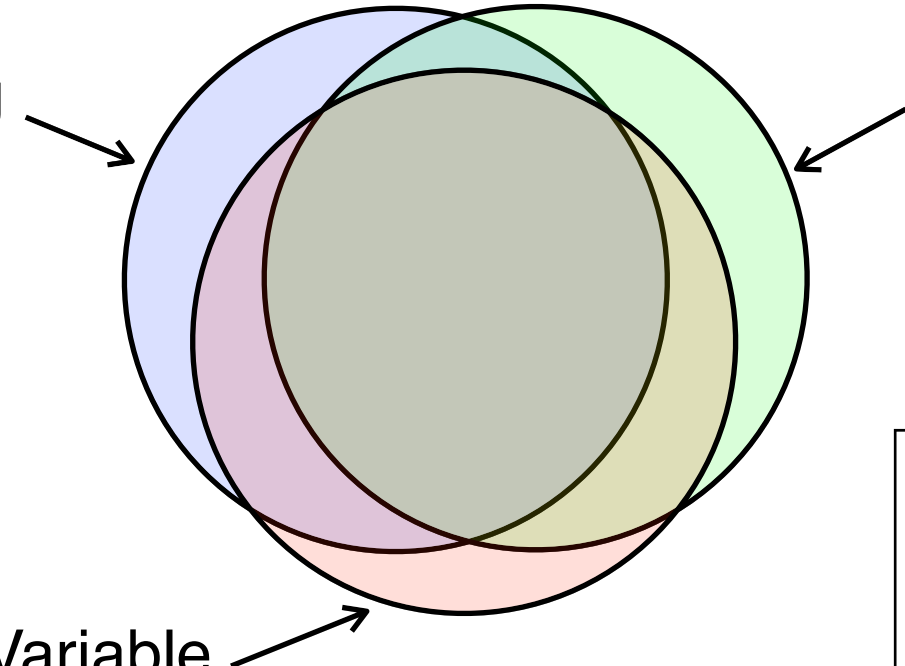
However, in an equally plausible, reverse interpretation of cause and effect: the association could come about because bad kids are not allowed to have dogs.

- Maybe smokers are on average more tense than non-smokers, and it's tension that disposes one toward getting cancer.
- Maybe smokers tend to drink a lot of coffee when smoking, and it's coffee that causes cancer.
- Maybe it's just that men happen to smoke more than women, and men also happen to be more vulnerable to lung cancer.

The Case of the Third Variable

Smoking

Lung Cancer



Third Variable

- Stress
- Coffee
- Gender

But even with the motivation and resources to meet each criticism, this process would be foiled by variables for which no good data were available (e.g., nervous tension).

Many of these can be rebutted by showing that controlling for them doesn’t eliminate the relationship between smoking and cancer.

For example, gender is a totally insufficient explanatory variable: Cancer rates are substantially higher for smokers than non-smokers, within *both* male and female populations.

The Case of the Third Variable

A better strategy is to spell out the details of the proposed causal mechanism, and then test the consequences...

Mechanism: Tobacco smoke contains substances that are toxic to human tissue when deposited by contact. The more contact, the more toxicity.

Now what are some empirical implications of such a mechanism?

1. The longer the person has smoked cigarettes, the greater the likelihood of cancer.
2. The more cigarettes a person smokes over a given period, the greater the likelihood of cancer.
3. People who stop smoking have lower cancer rates than those who keep smoking.
4. Smokers' cancer tend to occur in the lungs, and to be of a particular type.
5. Smokers have elevated rates of other respiratory diseases.
6. People who smoke cigars or pipes (where smoke isn't inhaled) have abnormally high rates of lip cancer.
7. Smokers of filter-tipped cigarettes have somewhat lower cancer rates than do other cigarette smokers.
8. Non-smokers who live with smokers have elevated cancer rates (presumably by passive exposure to smoke).

The Case of the Third Variable

All of these implications have moderate to strong empirical support and were established correlationally (by comparing cancer rates in different population subgroups).

Yet the case is extremely persuasive because it's so coherent. Furthermore, no additional explanatory mechanism seems required, as there are no anomalous results to be explained. If smokers were found to have four times the rate of nearsightedness, then this could create a nagging bit of incoherence, and keep the search open to new ideas.

A tight bundle of strong, plausible correlational results can be causally compelling. We can call this rebuttal strategy the *method of signatures*.

Kurt Cobain's troubled last days

Drugs, guns and threats; and then he disappeared

Phillips (1977) claimed a systematic connection between the dates of widely publicised suicides, and the number of motor vehicle accidents within the 7 day periods following these particular dates.

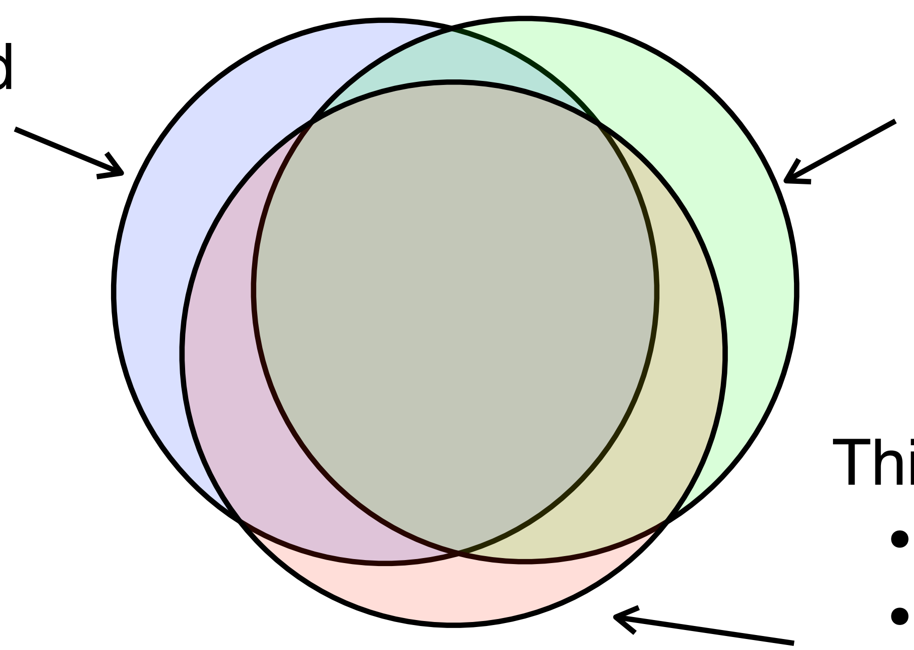
Mechanism: Publicised suicides encourage people with suicidal inclinations to take self-destructive action, one form of which is to deliberately crash a car.

But we should be especially suspicious of correlations between variables over time, because all kinds of events that have nothing to do with each other can co-occur in yearly, monthly, or weekly synchrony:

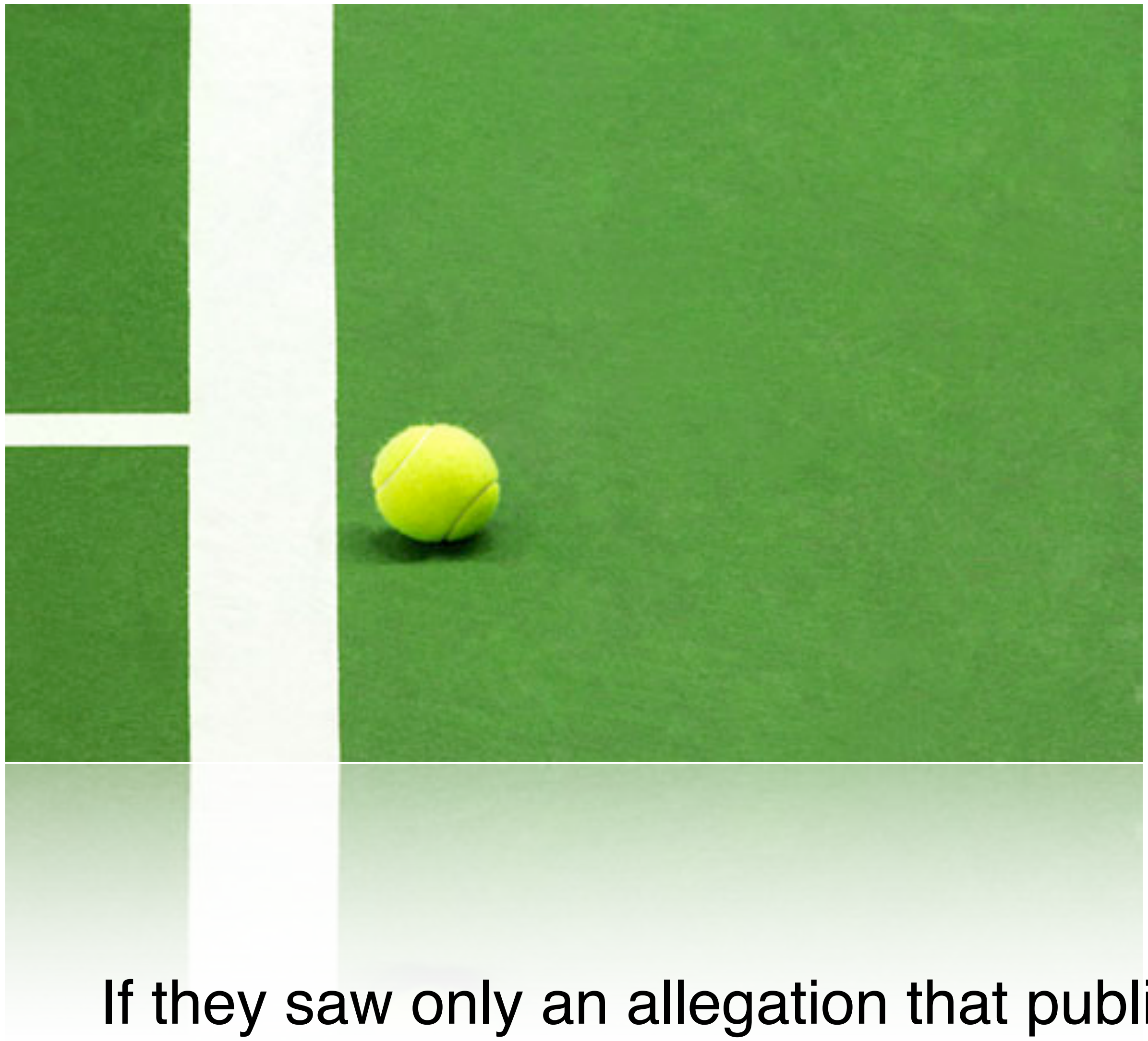
- Leap years
- Elections
- Betting on sport

Publicised suicides

Motor vehicle accidents



Third Variable
- Certain days of the week have more suicides?
- Holiday weekends?
- Any other national/international crisis (e.g., war, terrorism, stock market crash) may result in mass stress, worse driving, and more suicides.



Burden of Proof

Using a tennis metaphor, the toughest critics wouldn't even acknowledge that the ball was in their court.

If they saw only an allegation that publicised suicides were systematically followed by traffic accidents, they would call the researcher's shot out of bounds, and not respond until the opponent produced a better serve.

The investigator would be better off presenting a *signature* – a bundle of evidence consistent with the hypothesis, and inconsistent with other explanations.

For example, Phillips (1986) found that suicides that received heavier publicity were followed by more automobile fatalities and fatal traffic accidents tended to be confined to cases with a lone driver.

These results begin to fill in a signature characterising a genuine link.

Important concepts so far...

- The importance of linear composites in multivariate analysis
  - a linear composite is a weighted ‘average’ of the variables
  - forming a linear composite reduces many variables to one
- Variances (and sums of squares) can be partitioned

Important concepts so far...

- Correlation is the basis for multivariate methods.
- One view of data analysis is that we are trying to model our data by using linear composites
  - Residuals give information on the lack of fit between model and data