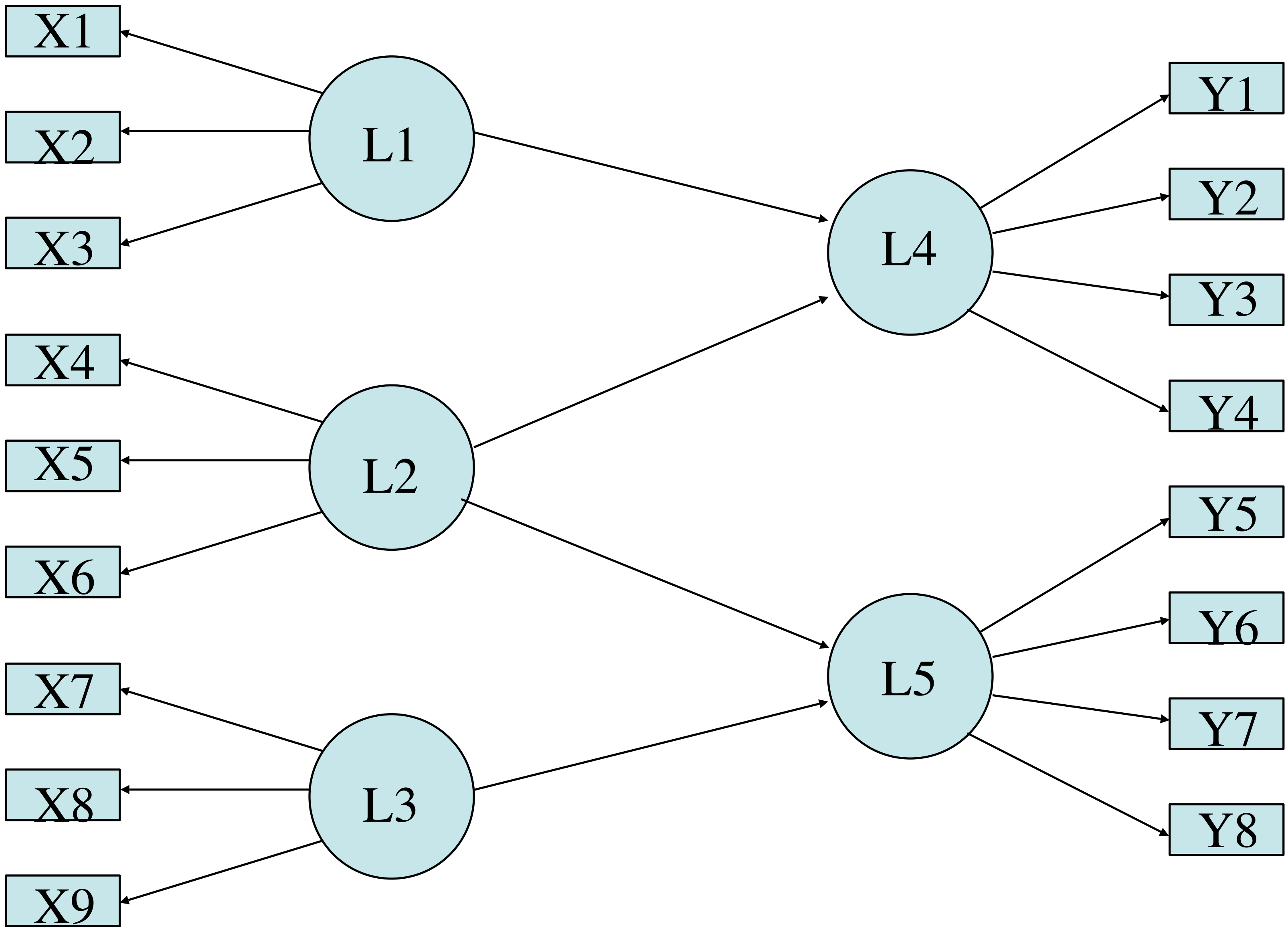
Psychometric Theory: A conceptual Syllabus



Psychometric Theory

Basic Concepts of Variance, Covariance and Correlation

Basic statistics

- Central tendency
  - multiple measures, multiple ways of measuring
- Measures of dispersion
  - Single variables
  - composite variables
- Measures of relationship
  - Bivariate
  - Multivariate

Estimates of Central Tendency

- Consider a set of observations $X = \{x_1,x_2\ldots x_n\}$
- What is the best way to characterize this set
  - Mode: most frequent observation
  - Median: middle of ranked observations

Mean:

$$\text{Arithmetic} = \overline{X} = \sum_{1}^{n}(X_i)/N$$

$$\text{Geometric} = \sqrt[n]{\prod_{1}^{n}(X_i)}$$

$$\text{Harmonic} = \frac{N}{\sum_{1}^{n}(1/X_i)}$$

Alternative expressions of mean

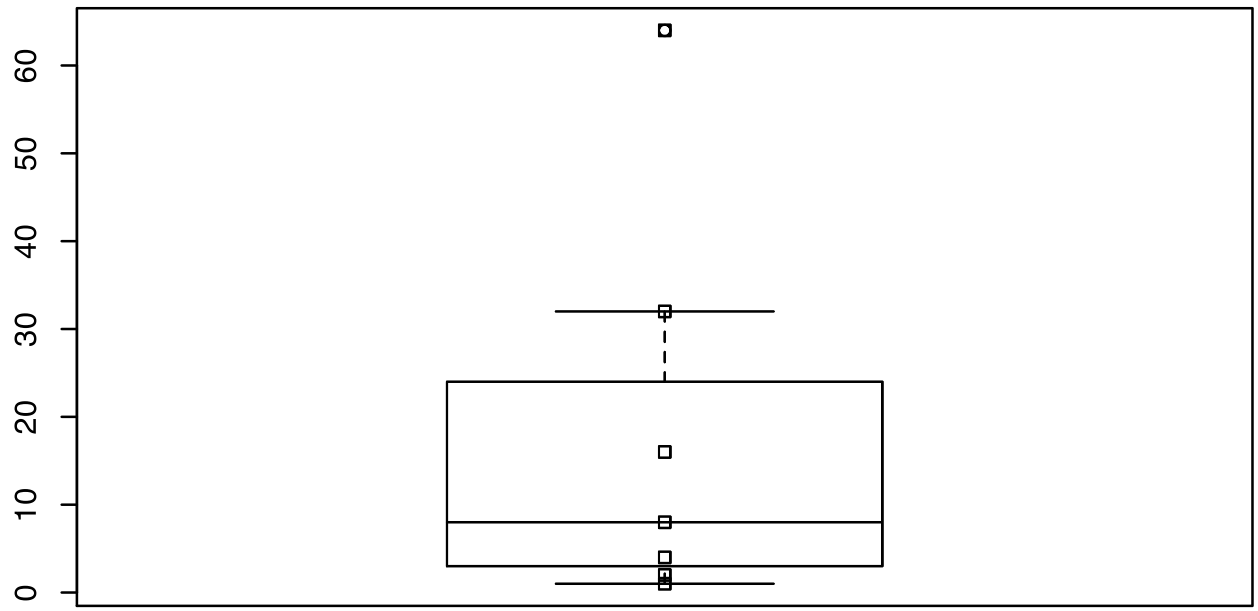
- Arithmetic mean = $\sum x_i/N$
- Alternatives are anti transformed means of transformed numbers
- Geometric mean = $\exp(\sum \ln(x_i)/N)$
  - (anti log of average log)
- Harmonic Mean = reciprocal of average reciprocal
  - $1/(\sum (1/x_i)/N)$

Why all the fuss?

- Consider 1,2,4,8,16,32,64
- Median = 8
- Arithmetic mean = 18.1
- Geometric = 8
- Harmonic = 3.5
- Which of these best captures the “average” value?

Summary stats (R code)

```
> x<-c(1,2,4,8,16,32,64)  #enter the data
> summary(x)  # simple summary
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    3.00    8.00   18.14   24.00   64.00
> boxplot(x)  #show five number summary
> stripchart(x,vertical=T,add=T)  #add in the points
```



Consider two sets, which is more?

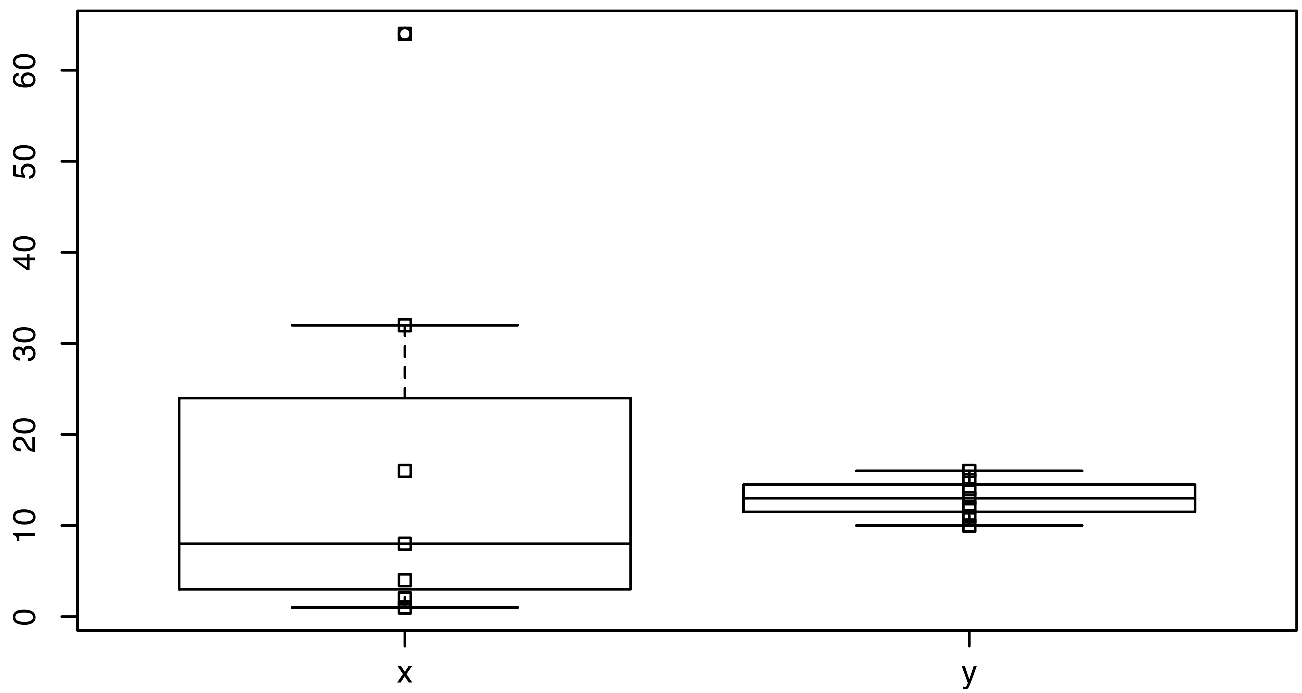
	subject	Set 1	Set 2
	1	1	10
	2	2	11
	3	4	12
	4	8	13
	5	16	14
	6	32	15
	7	64	16
median		8	13
arithmetic		18.1	13.0
geometric		8.0	12.8
harmonic		3.5	12.7

Summary stats (R code)

```
> x <- c(1,2,4,8,16,32,64)  #enter the data
> y <- seq(10,16)   #sequence of numbers from 10 to 16
> xy.df <- data.frame(x,y)  #create a "data frame"
> xy.df      #show the data
   x  y
1  1 10
2  2 11
3  4 12
4  8 13
5 16 14
6 32 15
7 64 16
> summary(xy.df)   #basic descriptive stats
       x               y       
 Min.   : 1.00   Min.   :10.0  
 1st Qu.: 3.00   1st Qu.:11.5  
 Median : 8.00   Median :13.0  
 Mean   :18.14   Mean   :13.0  
 3rd Qu.:24.00   3rd Qu.:14.5  
 Max.   :64.00   Max.   :16.0  
```

Box Plot (R)

boxplot(xy.df) #show five number summary
stripchart(xy.df,vertical=T,add=T) #add in the points



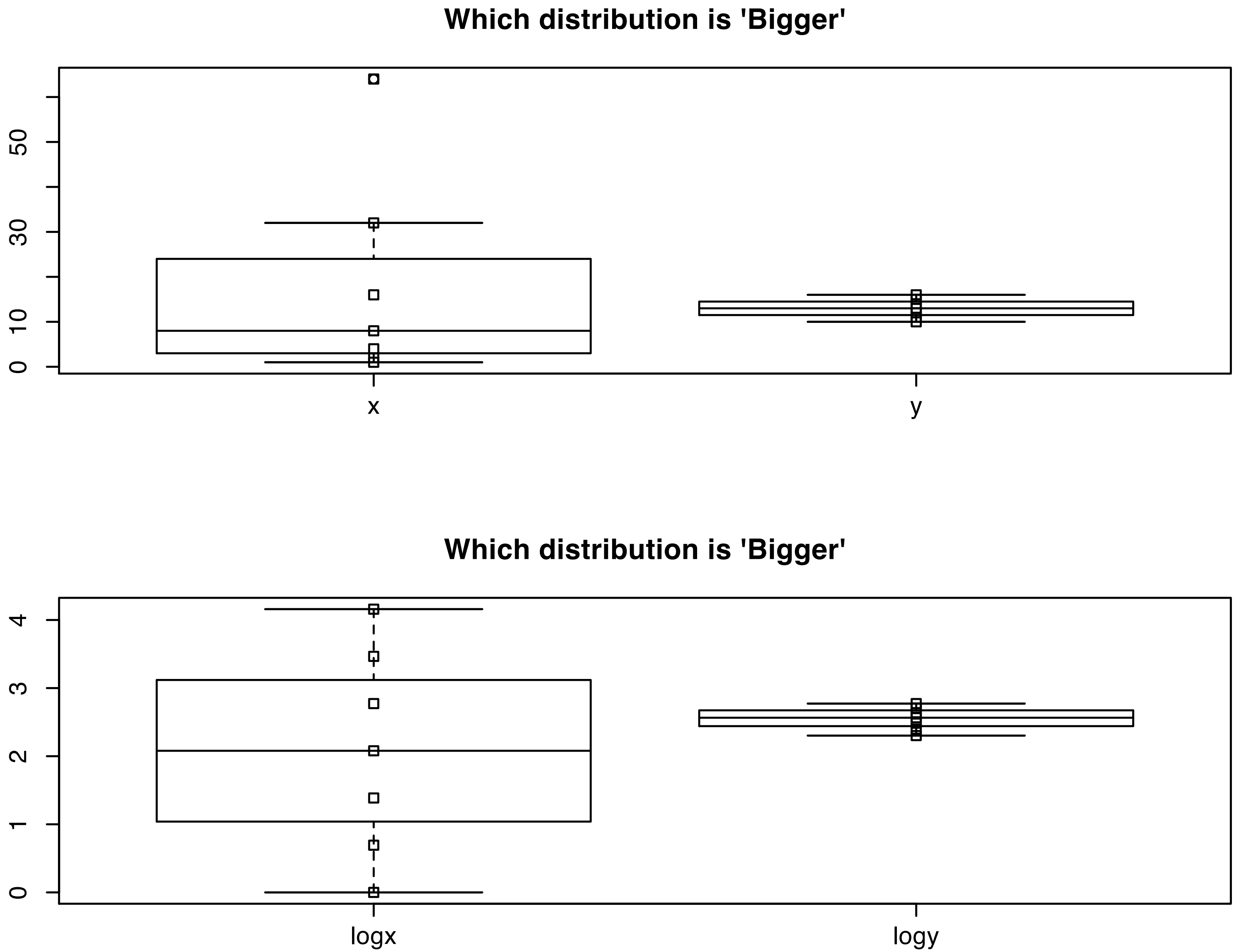
The effect of log transforms Which group is “more”?

X	Y	Log X	Log Y
1	10	0.0	2.3
2	11	0.7	2.4
4	12	1.4	2.5
8	13	2.1	2.6
16	14	2.8	2.9
32	15	3.5	2.7
64	16	4.2	2.8

Raw and log transformed which group is “bigger”?

	X	Y	Log(X)	Log(Y)
Min	1	10	0	2.30
1st Q.	3	11.5	1.04	2.44
Median	8	13	2.08	2.57
Mean	18.1	13	2.08	2.26
3rd Q.	24	14.5	3.12	2.67
Max	64	16	4.16	2.77

The effect of a transform on means and medians



Estimating central tendencies

- Although it seems easy to find a mean (or even a median) of a distribution, it is necessary to consider what is the distribution of interest.
- Consider the problems of the average length of psychotherapy, the average size of a class at NU, or the average velocity of cars on a highway.

Estimating the mean time of therapy

- A therapist has 20 patients, 19 of whom have been in therapy for 26-104 weeks (median, 52 weeks), 1 of whom has just had their first appointment. Assuming this is her typical load, what is the average time patients are in therapy?
- Is this the average for this therapist the same as the average for the patients seeking therapy?

Estimating the mean time of therapy

- 19 with average of 52 weeks, 1 for 1 week
  – Therapists average is (19*52+1*1)/20 = 49.5 weeks
  – Median is 52
- But therapist sees 19 for 52 weeks and 52 for one week so the average length is
  – ((19*52)+(52*1))/(19+52) = 14.6 weeks
  – Median is 1

Estimating Class size

5 faculty members teach 20 courses with the following distribution: What is the average class size?

Faculty member	100 fr	200 so-jr	300 jr-sr	400 grad	average
1	10	20	10	10	12.5
2	10	20	10	10	12.5
3	10	20	10	10	12.5
4	100	20	20	10	37.5
5	400	100	100	100	175
department	106	36	30	28	50

Estimating class size

- What is the average class size?
- If each student takes 4 courses, what is the average class size from the students' point of view?
- Department point of view: average is 50 students/class

N	Size
10	10
5	20
4	100
1	400

Estimating Class size

Faculty member	A	B	C	D	average
1	10	20	10	10	12.5
2	10	20	10	10	12.5
3	10	20	10	10	12.5
4	100	20	20	10	37.5
5	400	100	100	100	175
department	106	36	30	28	50

Estimating Class size (student weighted)

Faculty member	A	B	C	D	average
1	10	20	10	10	14
2	10	20	10	10	14
3	10	20	10	10	14
4	100	20	20	10	73
5	400	100	100	100	271
Student	321	64	71	74	203

Estimating class size

Department perspective:

20 courses, 1000 students => average = 50

Student perspective: 1000 students enroll in classes with an average size of 203!

Faculty perspective: chair tells prospective faculty members that median faculty course size is 12.5, tells the dean that the average is 50 and tells parents that most upper division courses are small.

Traffic Flow

- Three lanes of traffic, uniformly distributed
  - one lane is traveling at 10 mph
  - one lane is travelling at 20 mph
  - one lane is traveling at 30 mph
- What is the average velocity of cars?
- What is the median velocity?

Traffic Flow: But officer, I wasn't speeding

- Three lanes of traffic, uniformly distributed
  - one lane is traveling at 10 mph
  - one lane is travelling at 20 mph
  - one lane is traveling at 30 mph
- Assume cars are spaced a mile apart
  - Average = 30*30 +20*20 +10 *10 = 1400/60 =
    - 23.3
  - Median is 50th percentile -- mid point between 20 and 30 = 25

Average Velocity

- On a 100 mile trip from Chicago to Milwaukee, you drive the first 50 miles at 30 miles/hour and the second half at 60 miles/hour. What is your average velocity?
- A race car driver has to average 90 miles an hour for two laps of a one mile track. He does the first lap at 45 mph. How fast must he drive the second lap?

Velocity leads to time weighting

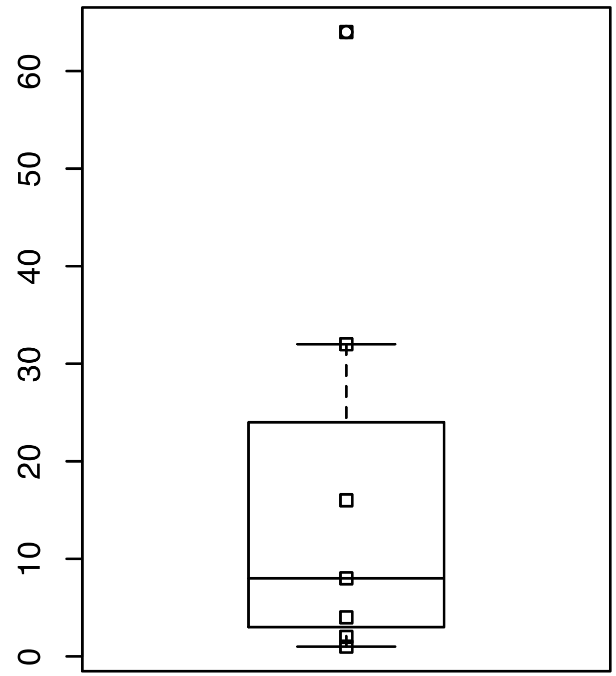
- A trip to Milwaukee:
  - 50 miles at 30 mph = 1.66 hours
  - 50 miles at 60 mph = .833 hours.
  - Average is (1.66* 30 + .833 * 60)/2.5 = 40 mph
- Race car driver
  - First lap at 45 => 1.33 minutes
  - Total time allowed = 120 secs/90 =1.33 minutes
  - driver can not average 90!

Measures of dispersion

- Range (maximum - minimum)
- Interquartile range (75% - 25%)
- Deviation score $x_i = X_i$-Mean
- Median absolute deviation from median
- Variance = $\sum x_i^2/(N-1)$ = mean square
- Standard deviation sqrt (variance) =sqrt ($\sum x_i^2/(N-1)$)

Robust measures of dispersion

- The 5-7 numbers of a box plot
- Max
- Top Whisker
- Top quartile (hinge)
- Median
- Bottom Quartile (hinge)
- Bottom Whisker
- Minimum



Raw scores, deviation scores and Standard Scores

- Raw score for i_{th} individual X_i
- Deviation score $x_i = X_i$-Mean X
- Standard score = x_i/s_x
- Variance of standard scores = 1
- Mean of standard scores = 0
- Standard scores are unit free index

Transformations of scores

- Mean of (X+C) = Mean(X) + C
- Variance (X+C) = Variance(X)
- Variance (X*C) = Variance(X) *C^2
- Coefficient of variation = sd/mean

Typical transformations

	Mean	Standard Deviation
Raw data	$X. = \sum X/n$	$Sqrt(\sum(X-X.)^2)/(n-1) = s_x = Sqrt(\sum x^2)/(n-1)$
deviation score	0	s_x
Standard score	0	1
"IQ"	100	15
"SAT"	500	100
"T-Score"	50	10
"stanine"	5	1.5

Variance of Composite

	X	Y
X	Variance_{X}	Covariance_{XY}
Y	Covariance_{XY}	Variance_{Y}

$$\text{Variance}_{(X+Y)} = \text{Var}_{X} + \text{Var}_{Y} + 2\ \text{Cov}_{XY}$$

Variance of Composite

	X	Y
X	$\sum x_i^2/(N-1)$	$\sum x_i y_i /(N-1)$
Y	$\sum x_i y_i /(N-1)$	$\sum$ yi 2/(N-1)

$$Var_{(X+Y)} = \sum(x_i+y_i)^2/(N-1) = \sum x_i^2/(N-1) + \sum y_i^2/(N-1) + 2 \sum x_i y_i /(N-1)$$

Consider the following problem

- If you have a GRE V of 700 and a GRE Q of 700, how many standard deviations are you above the mean GRE (V+Q)?
- Need to know the Mean and Variance of V, Q, and V+Q

	GRE V	GRE Q	GRE V+Q
Mean	500	500	1000
SD	100	100	?

Variance of GRE (V+Q)

	GRE V	GRE Q
GRE V	10,000	6,000
GRE Q	6,000	10,000

Variance of composite = 32,000 => s.d. composite = 179

Variance of GRE (V+Q)

	GRE_V	GRE_Q	GRE_{V+Q}
Mean	500	500	1000
SD	100	100	179

Standard score on composite

	GRE $_V$	GRE $_Q$	GRE $_{V+Q}$
mean	500	500	1000
sd	100	100	179
raw score	700	700	1400
z score	2	2	2.23
percentile	97.7	97.7	98.7

Variance of composite of n variables: generalization of variance of x+y

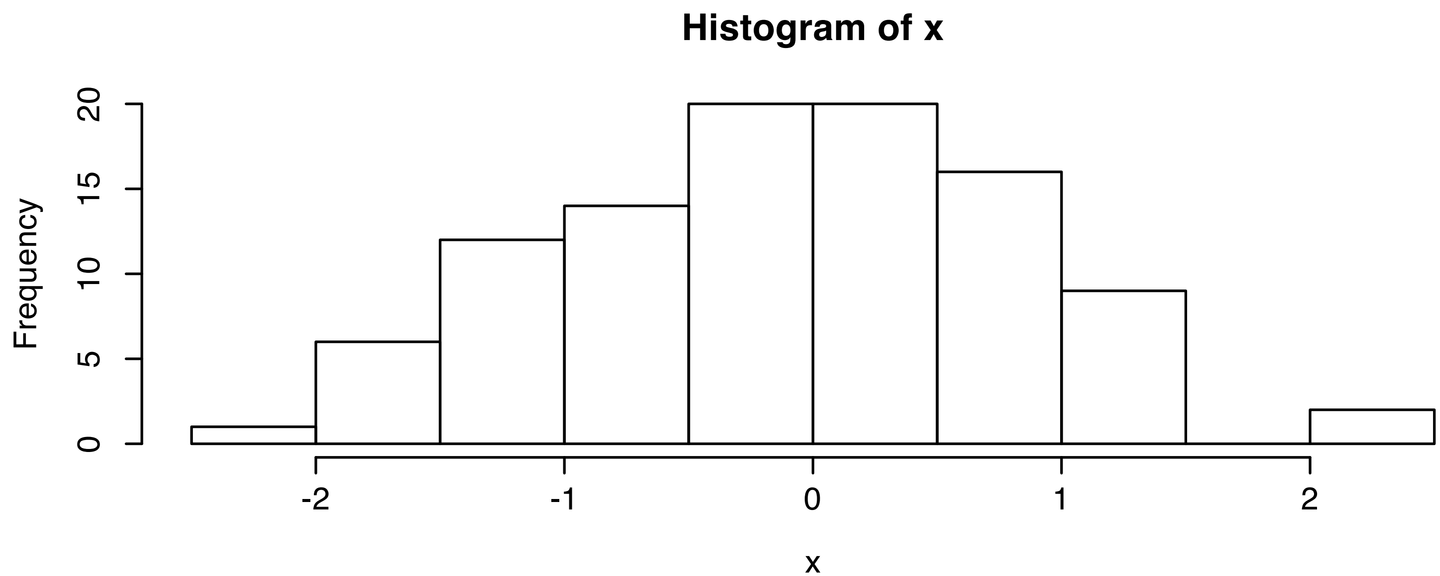
	X_1	X_2	…	X_i	X_j	…	X_n
X_1	Vx_1						
x_2	Cx_1x_2	Vx2					
…			…				
X_i	Cx_1x_i	Cx_2x_i		Vx_i			
X_j	Cx_1x_j	Cx_2x_j		Cx_ix_i	Vx_j		
…						…	
X_n	Cx_1x_n	Cx_2x_n		Cx_ix_n	Cx_jx_n		Vxn

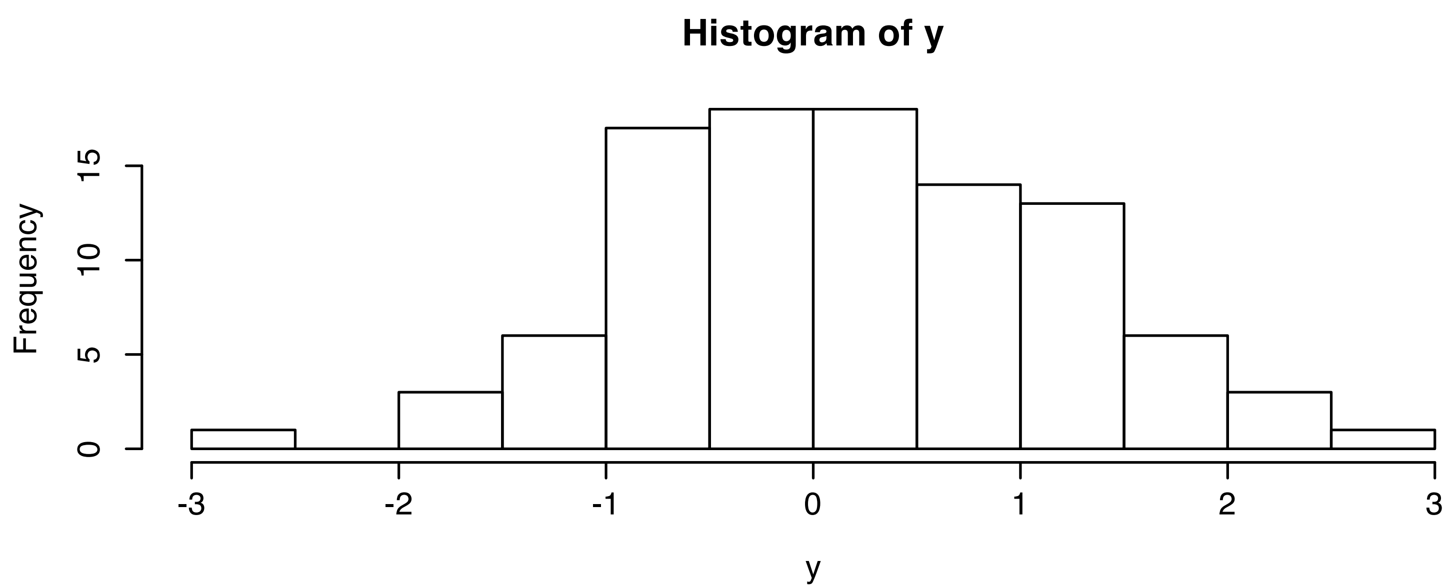
Variance of composite of n items has n variances and n*(n-1) covariances

Variance, Covariance, and Correlation

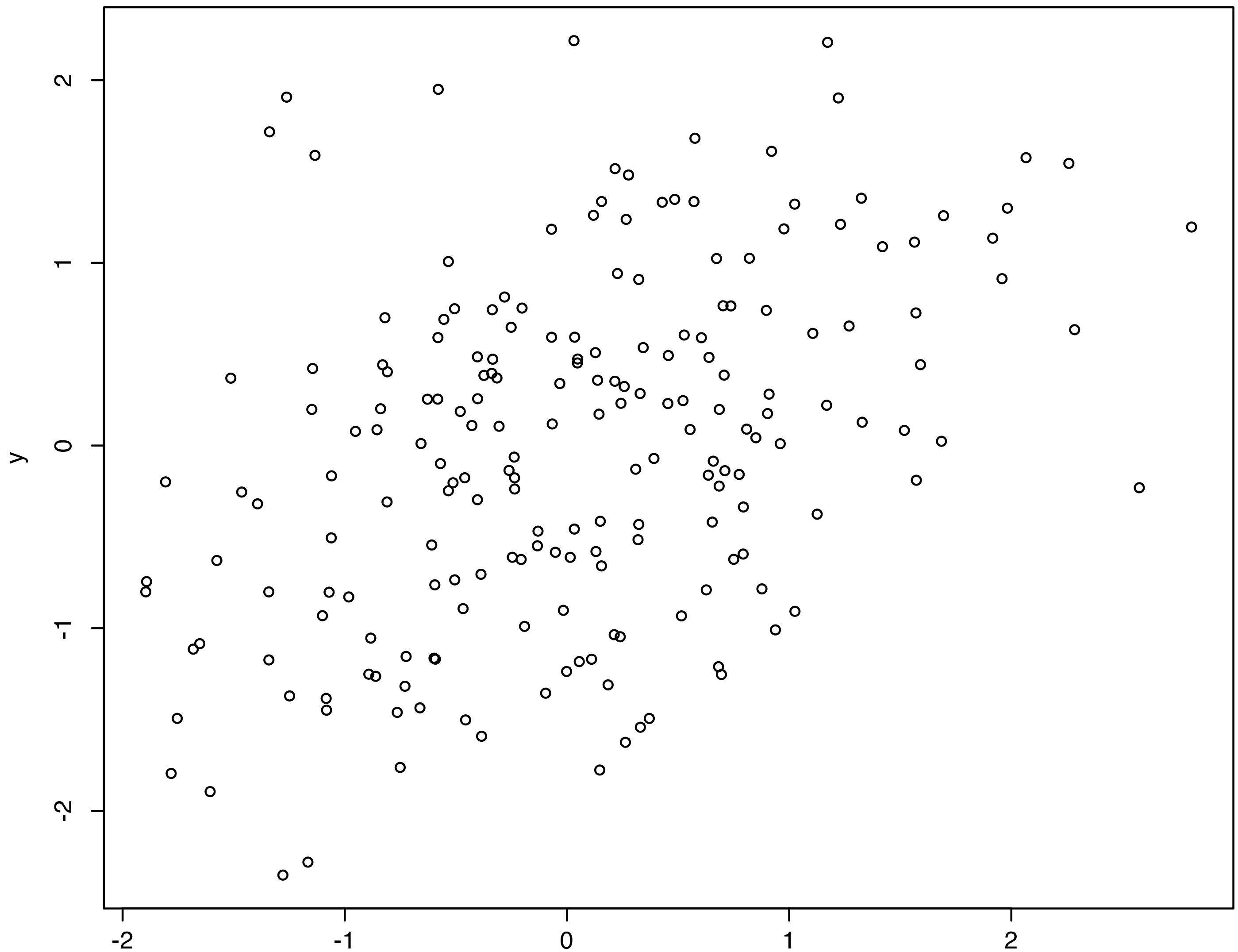
- Given two variables, X and Y, can we summarize how they interrelate?
- Given a score x_i, what does this tell us about y_i
- What is the amount of uncertainty in Y that is reduced if we know something about X.
- Example: the effect of daily temperature upon amount of energy consumed per day
- Example: the relationship between anxiety and depression

Distributions of two variables

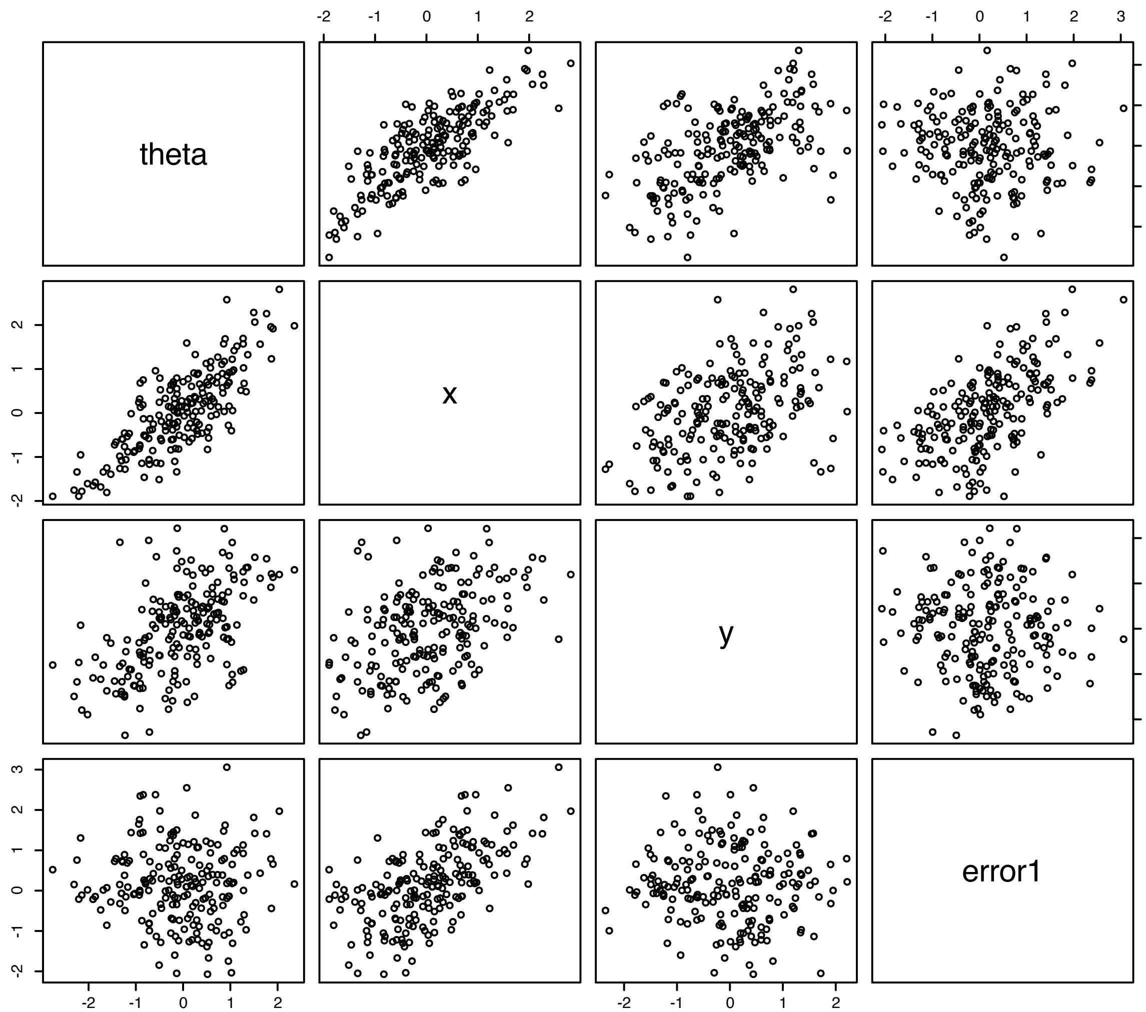




Joint distribution of X and Y



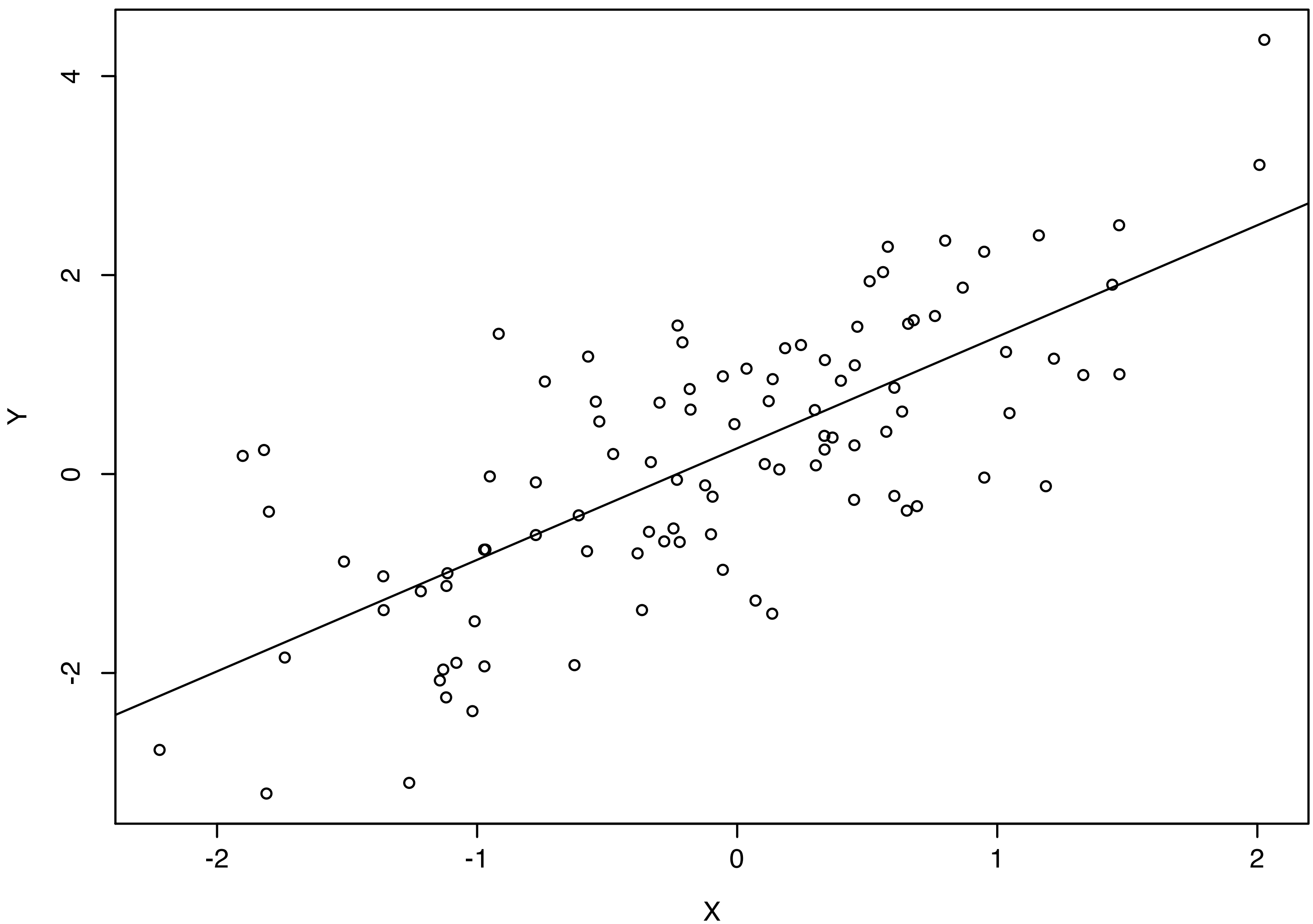
The problem of summarizing several bivariate relationships



Predicting Y from X

- First order approximation: predict mean Y for all y
- Second order approximation: predict y_i deviates from mean Y as linear function of deviations of x_i from mean X
- $Y_i = Y. + b_{xy}(X_i - X.)$ or $y_i = b_{xy}(x_i)$
- What is the best value of b_{xy}?

Predicting Y from X



The problem of predicting y from x:

- Linear prediction $y=bx+c$ $Y=b(X-M_x)+M_y$
- error in prediction = predicted y - observed y
- problem is to minimize the squared error of prediction
- minimize the error variance = $V_e = [\sum(y_p-y_o)^2]/(N-1)$
- $V_e=V_{(bx-y)} = \sum (bx-y)^2/(N-1)=$
- $\sum(b^2x^2-2bxy+y^2)/(N-1)=$
- $b^2\sum x^2/(N-1)-2b\sum xy/(N-1)+\sum y^2/(N-1)==>$
- $V_e = b^2V_x-2bC_{xy}+V_y$

- V_e is minimized when the first derivative (w.r.t. b) = 0 ==>
- when $2bV_x-2C_{xy}=0$ ==>
- $b_{y.x}=C_{xy}/V_x$

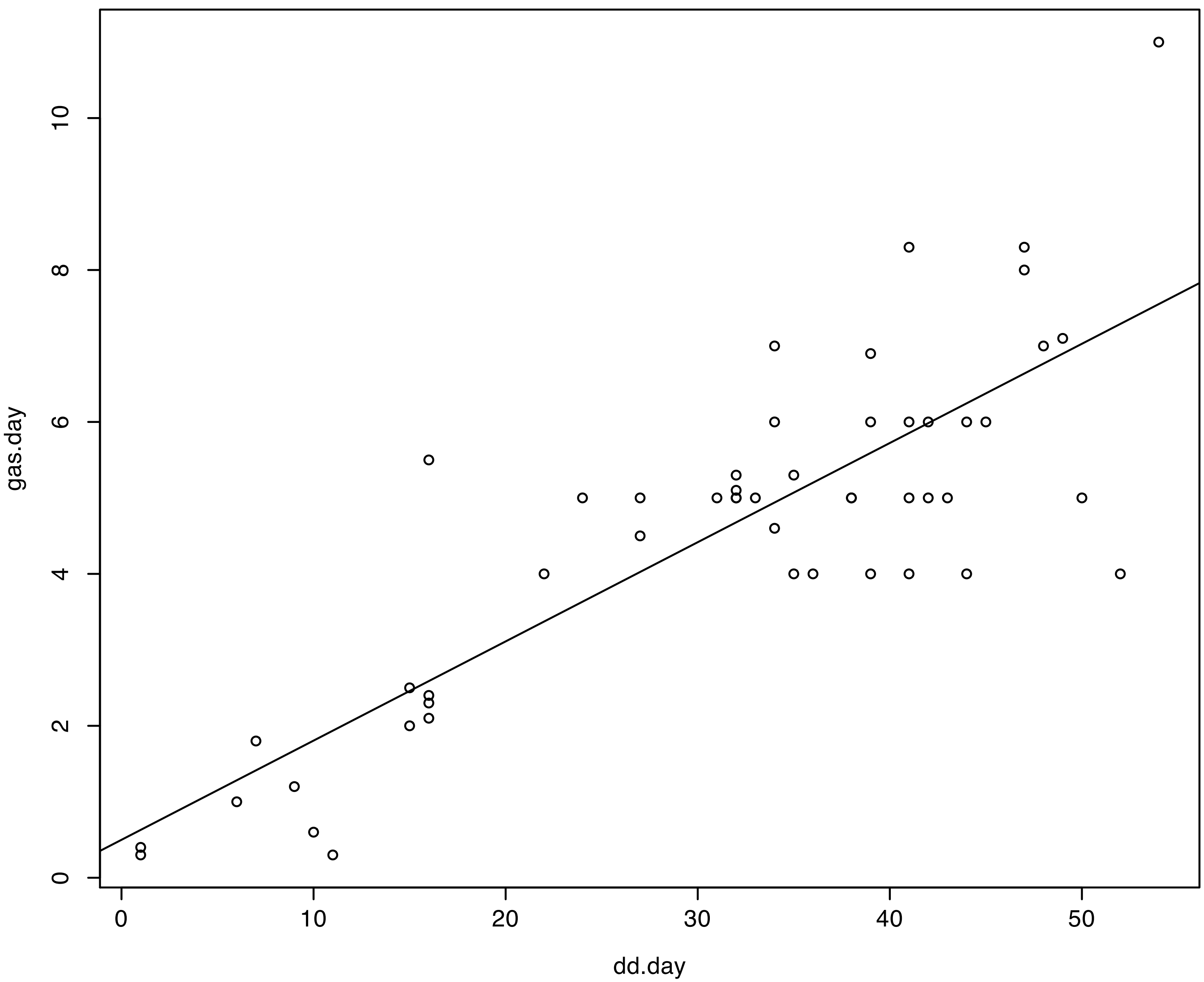
Measures of relationship

- Regression $y = bx + c$
  - $b_{y.x} = Cov_{xy} / Var_{x}$ $b_{x.y} = Cov_{xy} / Var_{y}$
- Correlation
  - $r_{xy} = Cov_{xy}/sqrt(V_x * V_y)$
  - Pearson Product moment correlation
    - Spearman (ppmc on ranks)
    - Point biserial (x is dichotomous, y continuous)
    - Phi (x, y both dichotomous)

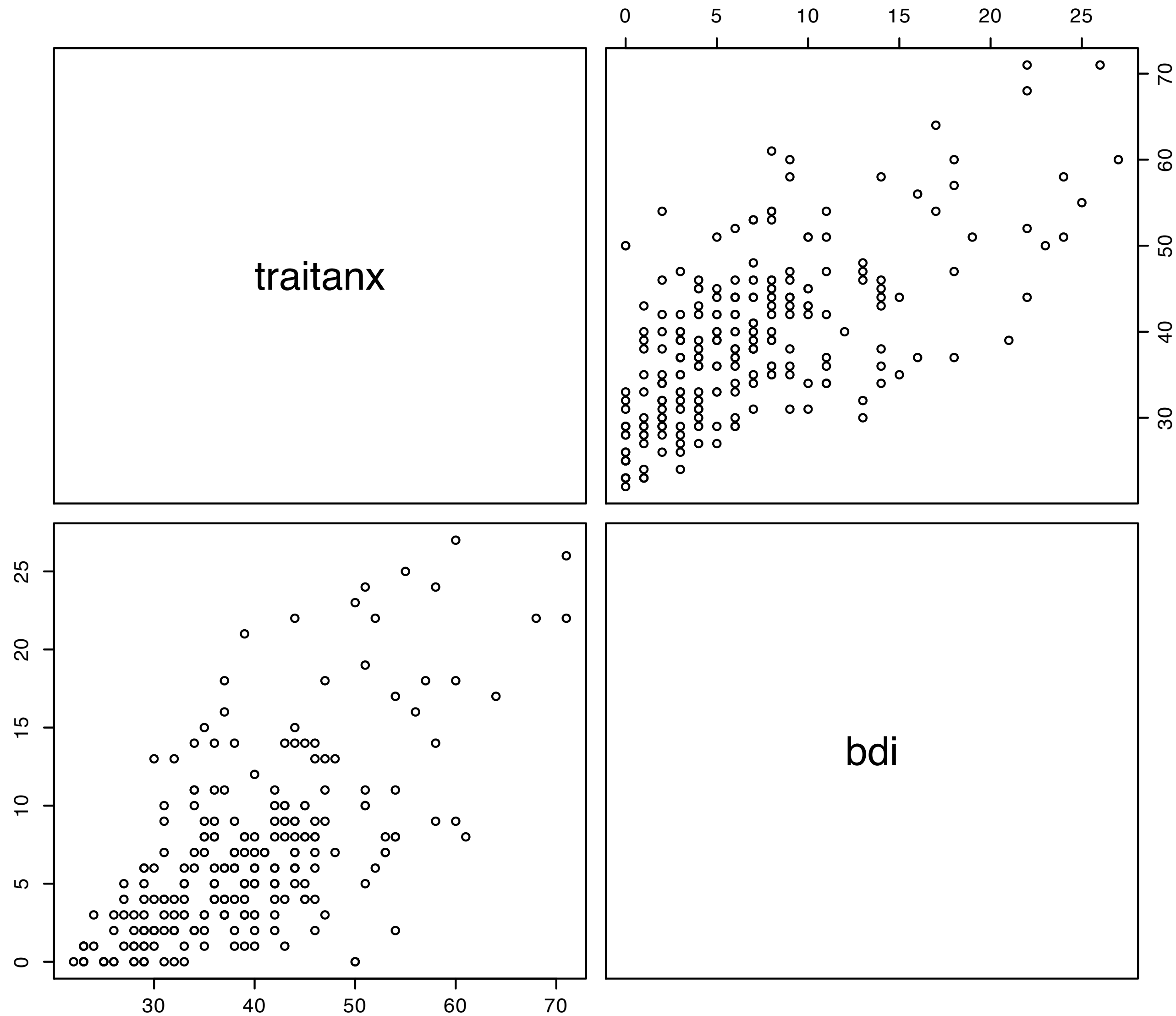
Correlation and Regression

- Regression slope is in units of DV and IV
  - regression implies IV -> DV
    - (gas consumption as function of outside temp)
- Correlation is unit free index of relationship
  - (geometric) average of two regression slopes
  - slope of standardized IV regression on standardized DV => unit free index
  - a measure of goodness of fit of regression

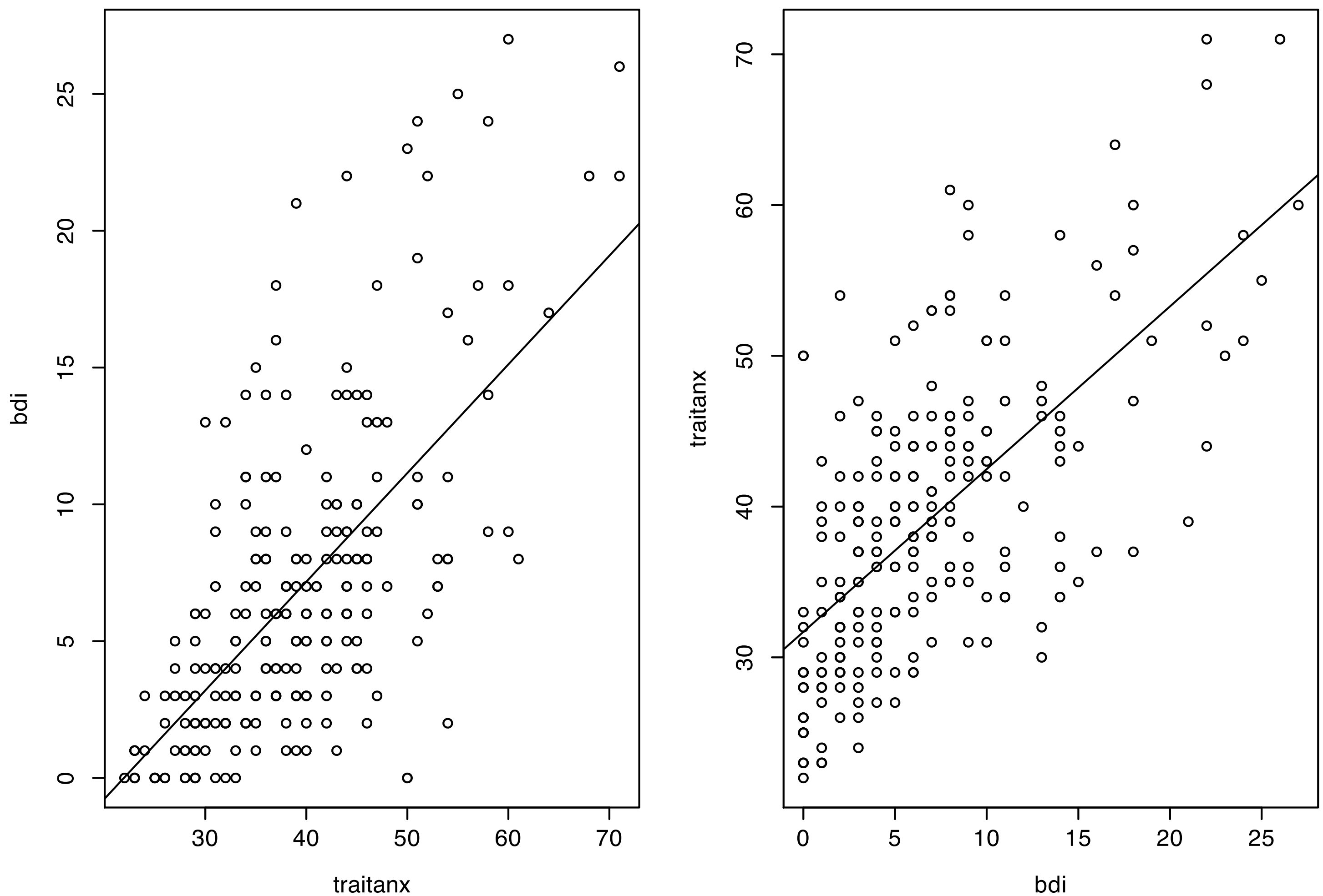
Gas Consumption by degree day (daily data)



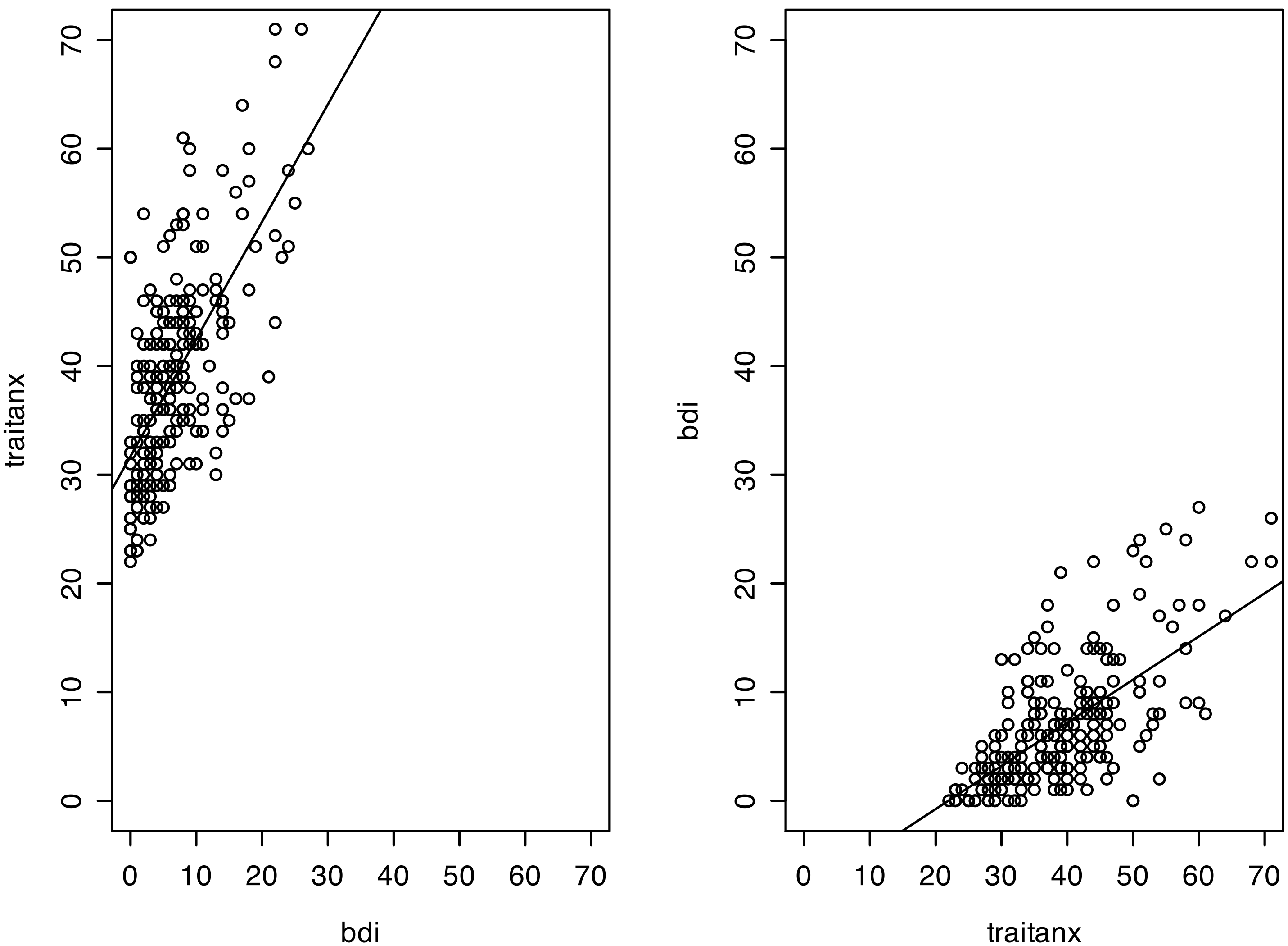
Beck Depresion x Trait Anxiety (raw)



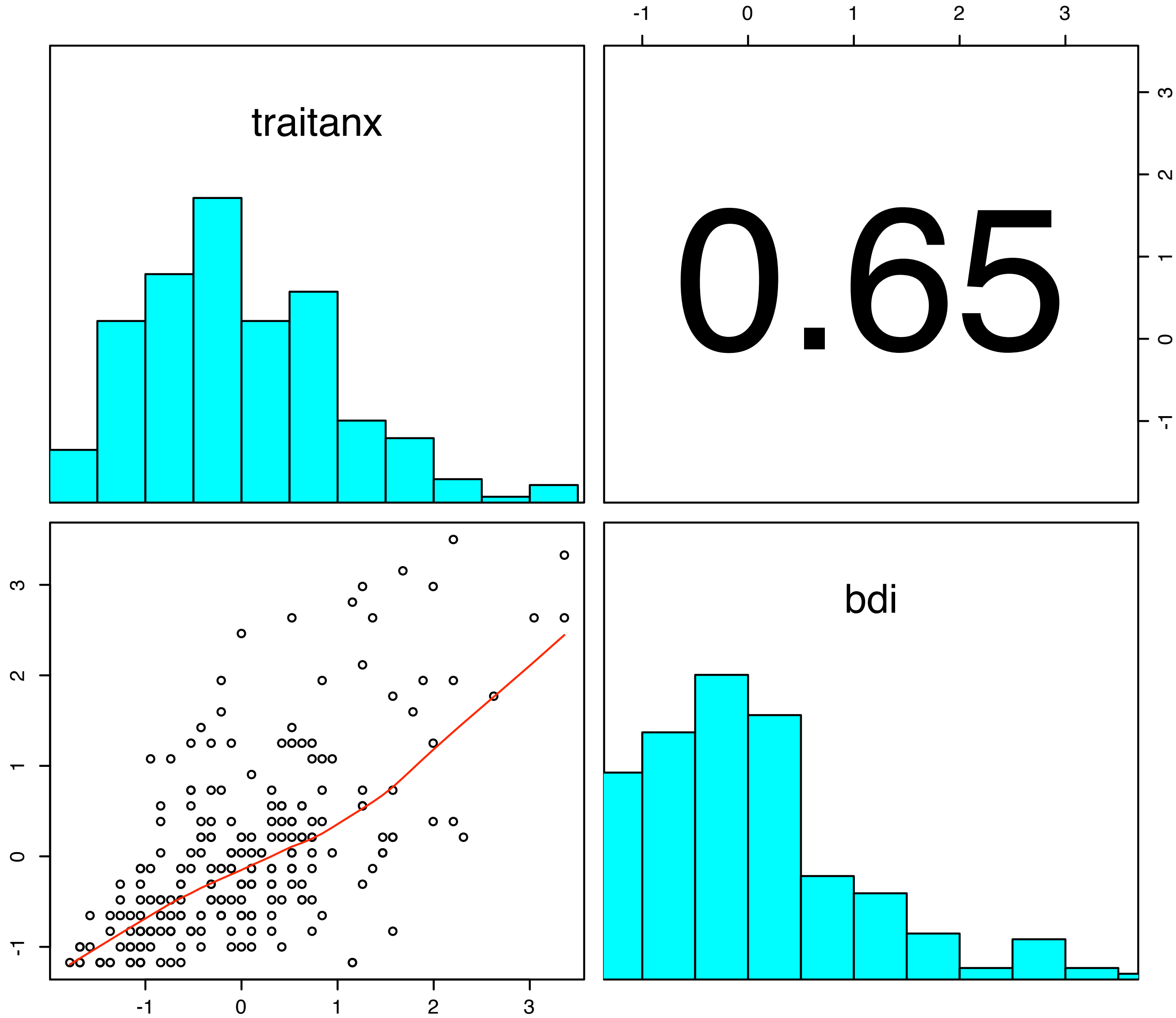
BDI x Trait Anx (raw)



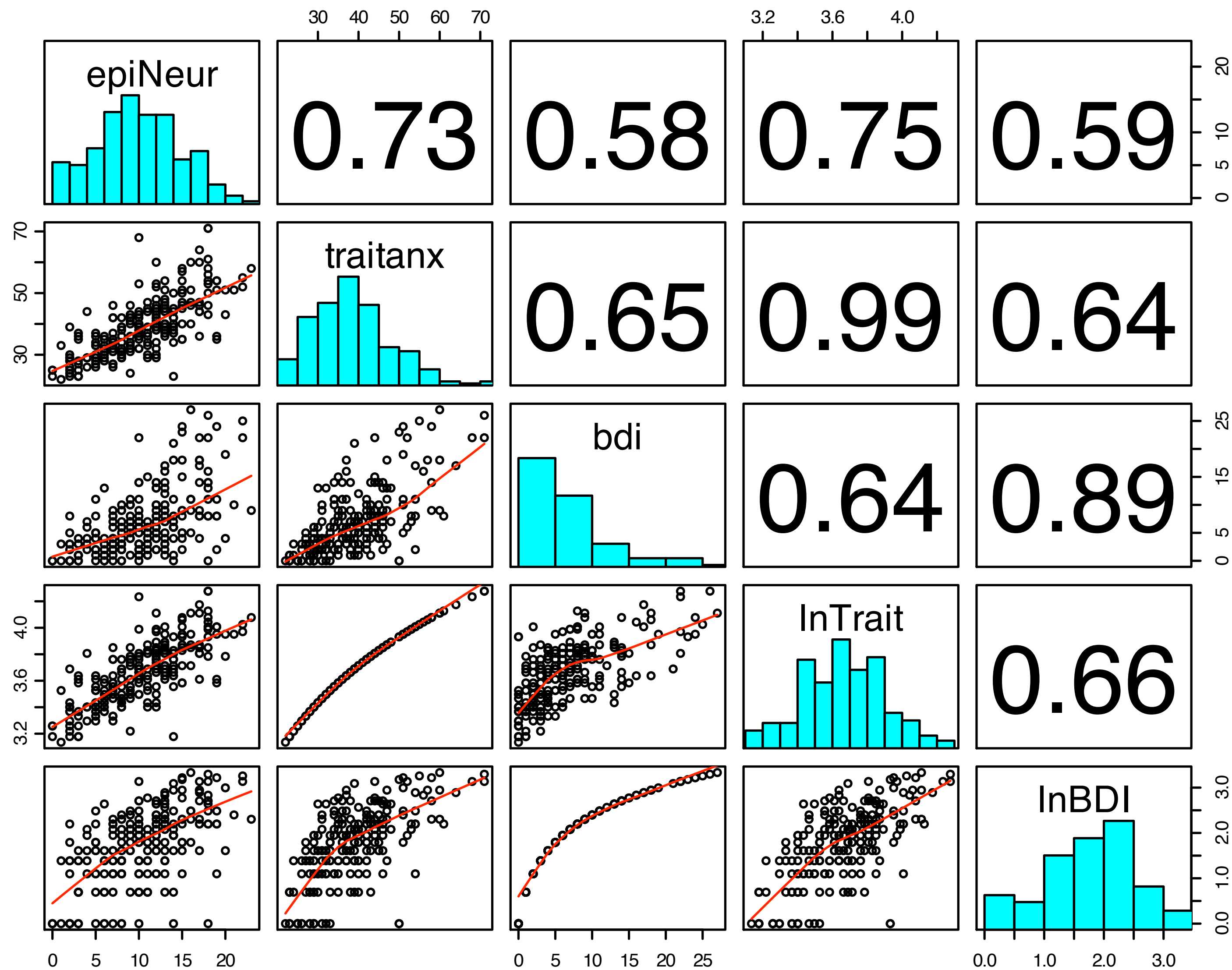
Regression lines depend upon scale



Beck Depression * Trait Anxiety z score



Transforming can help



Alternative forms of r

$r = cov_{xy}/Sqrt(V_x * V_y) =$

$(\sum xy/N)(sqrt(\sum x^2/N * \sum y^2/N) = (\sum xy)(sqrt(\sum x^2 * \sum y^2)$

Correlation	X	Y
Pearson	Continuous	Continuous
Spearman	Ranks	ranks
Point biserial	Dichotomous	Continuous
Phi	Dichotomous	Dichotomous
Biserial	Dichotomous (assumed normal)	Continuous
Tetrachoric	Dichotomous (assumed normal)	Dichotomous (assumed normal
Polychoric	categorical (assumed normal)	categorical (assumed normal)

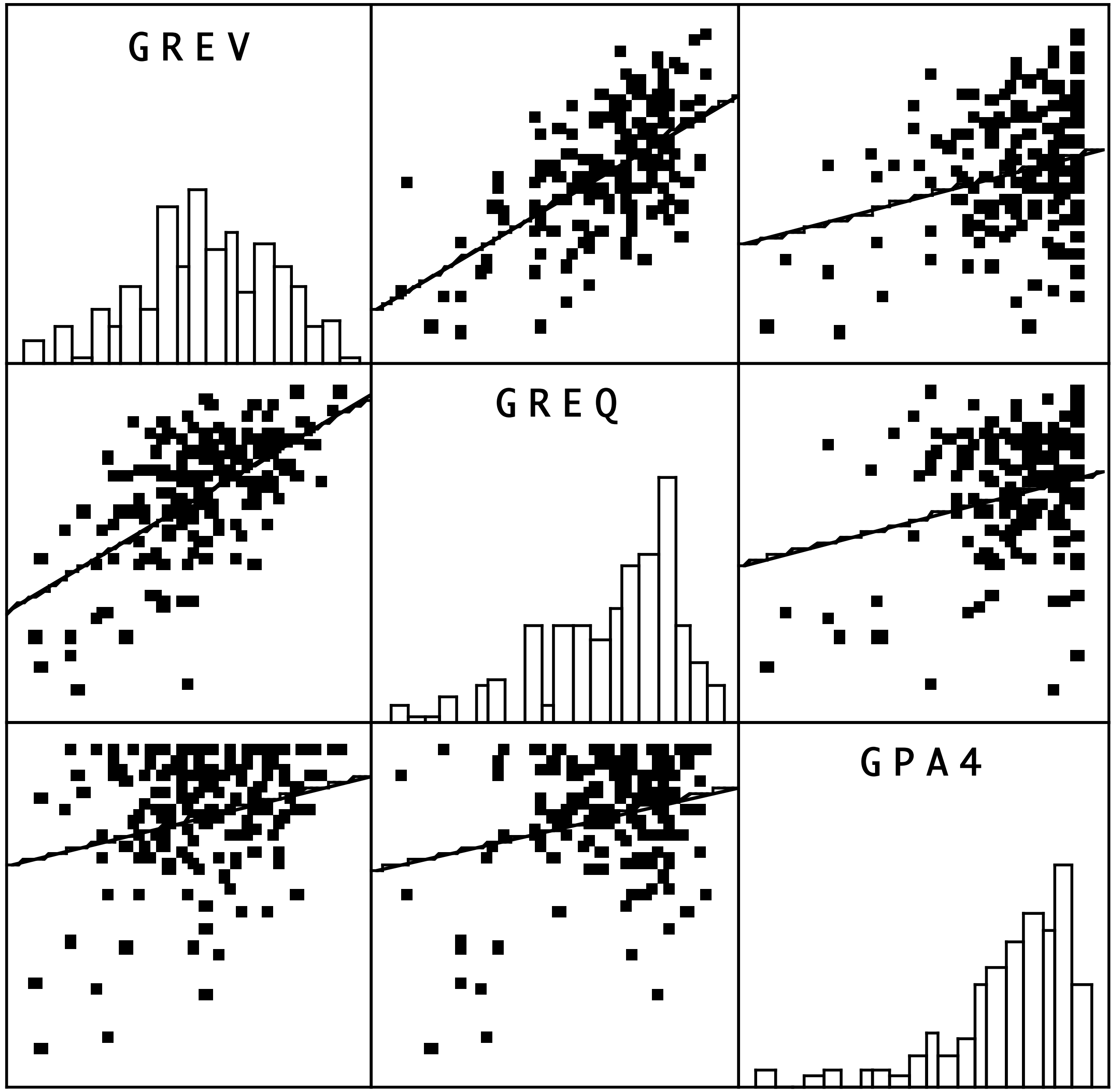
Correlation Matrix: GRE V, Q, GPA

PEARSON CORRELATION MATRIX

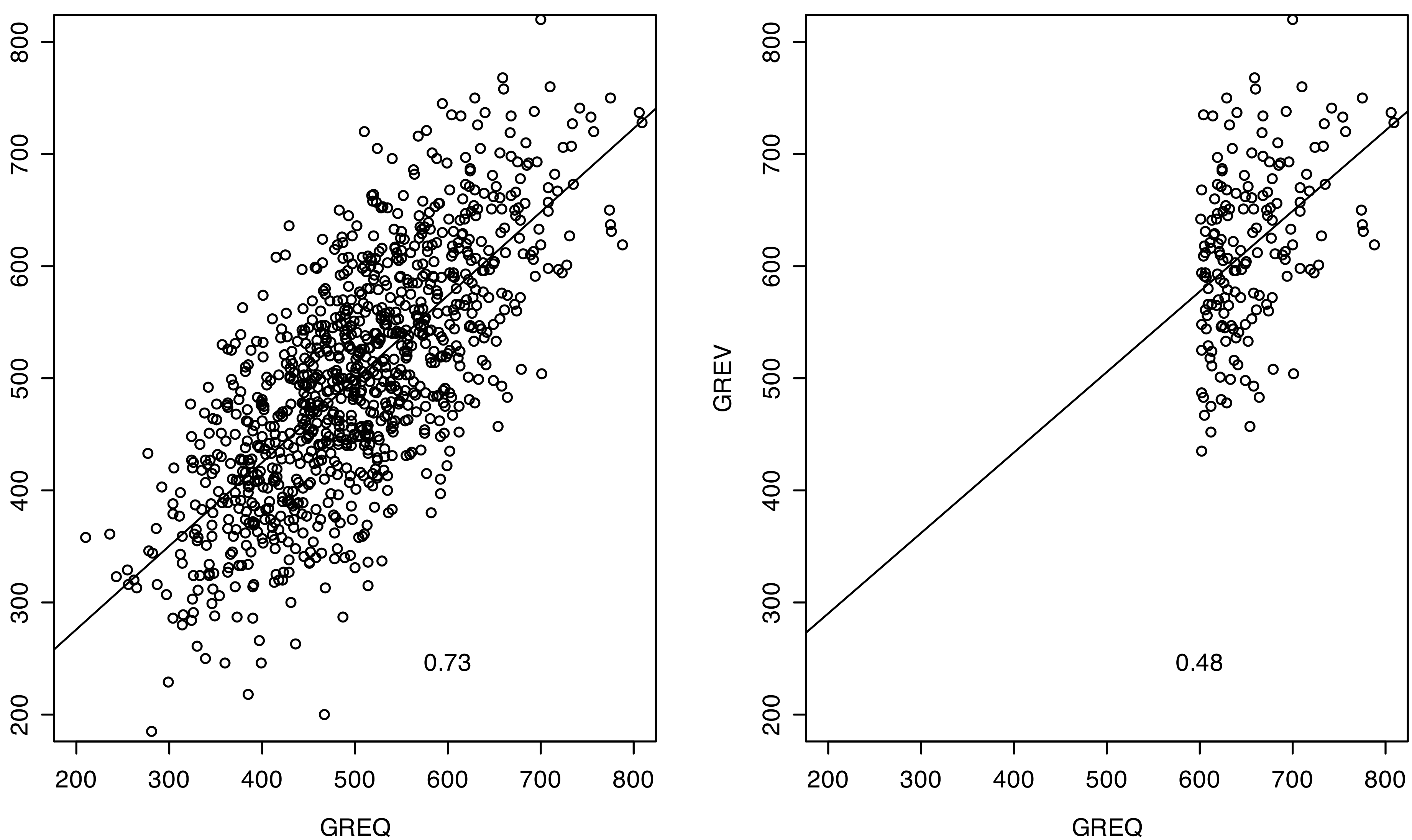
	GREV	GREQ	GPA4
GREV	1.00		
GREQ	0.61	1.00	
GPA4	0.27	0.25	1.00

NUMBER OF OBSERVATIONS: 163

SPLOM of GRE V, Q, GPA



The effect of restriction of range on regression slopes vs. correlations



Caution with correlation

Consider 8 variables with means:

x1	x2	x3	x4	y1	y2	y3	y4
9.0	9.0	9.0	9.0	7.5	7.5	7.5	7.5

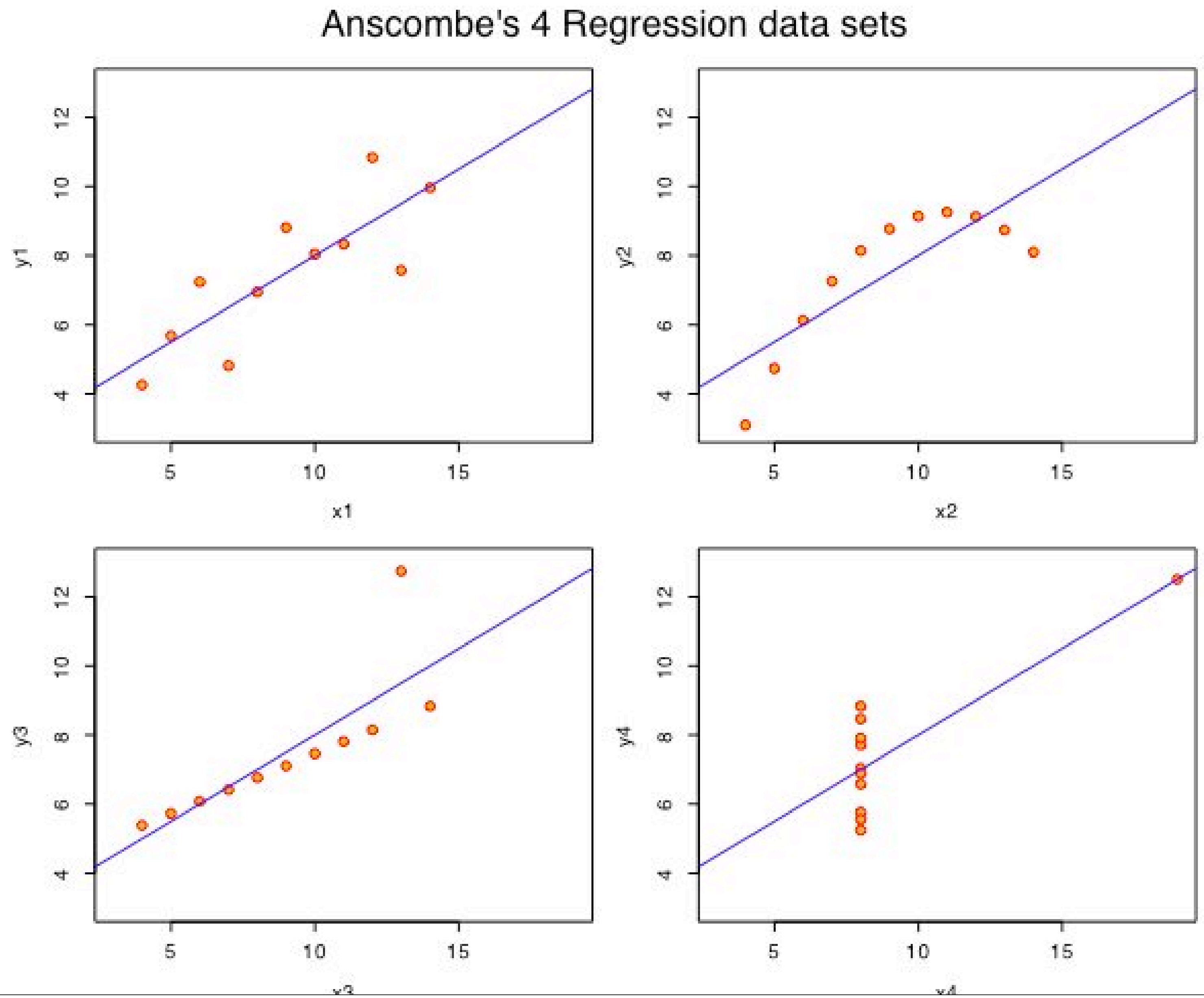
and Standard deviations

x1	x2	x3	x4	y1	y2	y3	y4
3.32	3.32	3.32	3.32	2.03	2.03	2.03	2.03

and correlations between xi and yi of

0.82 0.82 0.82 0.82

Caution with Correlation



Correlation: Alternative meanings

1) Slope of regression ($b_{xy} = C_{xy}/V_x$) reflects units of x and y but the correlation $\{r = C_{xy}/(S_xS_y)\}$ is unit free.

2) Geometrically, r = cosine (angle between test vectors)

3) Correlation as prediction:
Let y_p = predicted deviation score of y = predicted Y - M

$y_p = b_{xy}x$ and $b_{xy} = C_{xy}/V_x = rS_y/S_x$ ==> $y_p/S_y = r(x/S_x)$==>

predicted z score of y (z_{yp}) = r_{xy} * observed z score of x (z_x)

predicted z score of x (z_{xp}) = r_{xy} * observed z score of y (z_y)

Correlation as goodness of fit

Amount of error variance (residual or unexplained variance) in y given x and r

$V_e = \sum e^2/N = \sum y\text{-}bx)^2/N = \sum\{y-(r*S_y*x/S_x)\}^2 =$

$V_y + V_y * r^2 - 2(r * S_y * C_{xy})/S_x$

(but $S_y * C_{xy}/S_x = V_y * r$)

$V_y + V_y * r^2 - 2(r^2 * V_y) = V_y(1-r^2)$ ==>

$V_e = V_y(1-r^2)$ <==> $V_{yp} = V_y(r^2)$

Residual Variance = Original Variance * $(1-r^2)$

Variance of predicted scores = original variance * r^2

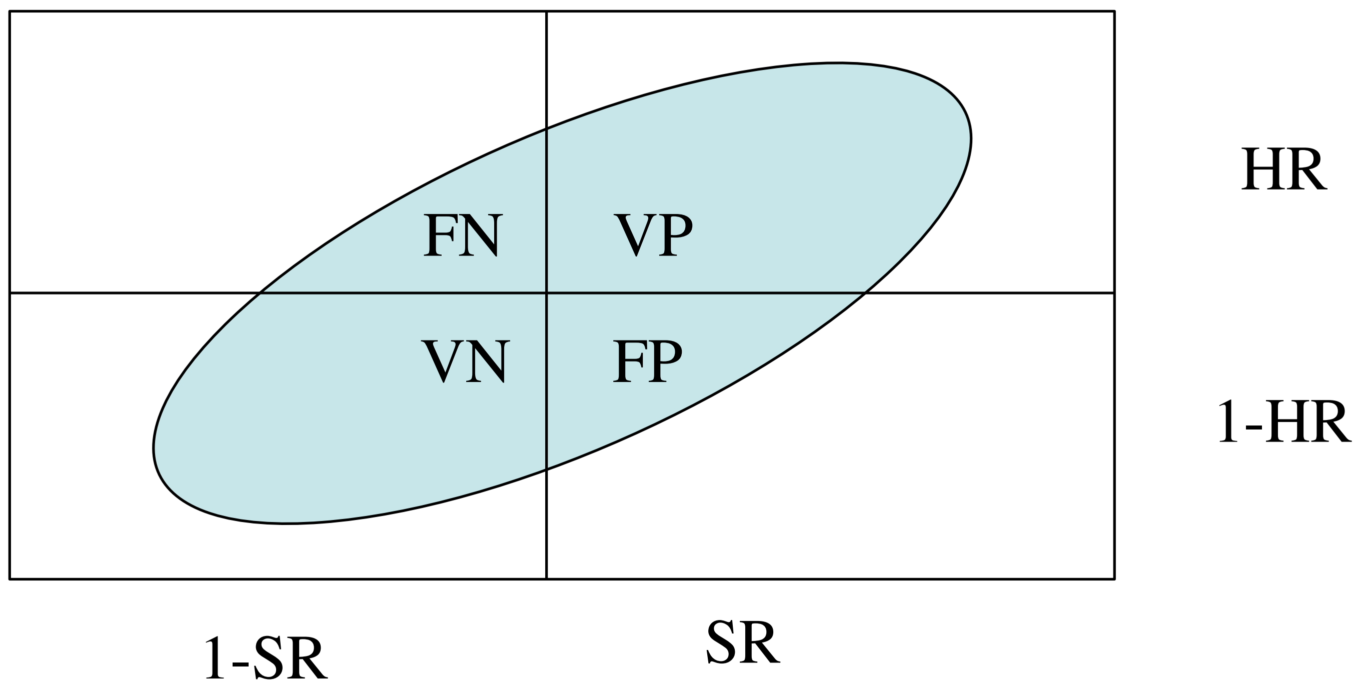
Basic relationships

	X	Y	Yp	Residual
Variance	V_x	V_y	$V_y(r^2)$	$V_y(1-r^2)$
Correl with X	1	r_{xy}	1	0
Correl with Y	r_{xy}	1	r_{xy}	$\sqrt{(1-r^2)}$

Phi coefficient of correlation

Hit Rate = Valid Positive + False Negative

Selection Ratio = Valid Positive + False Positive



Phi =(VP - HR*SR) /sqrt(HR*(1-HR)*(SR)*(1-SR)

Correlation size ≠ causal importance

	Pregnant	Not Pregnant	Total
Intercourse	2	1,041	1,043
No intercourse	0	6,257	6,257
Total	2	7,298	7,300

Correlation size ≠ causal importance

	Pregnant	Not Pregnant	Total
Intercourse	.0003	.1426	.1429
No intercourse	.0000	.8571	.8571
Total	.0003	.9997	1.0000

Phi =(VP - HR*SR) /sqrt(HR*(1-HR)*(SR)*(1-SR)= .04

polychoric rho = .53

Sex discrimination?

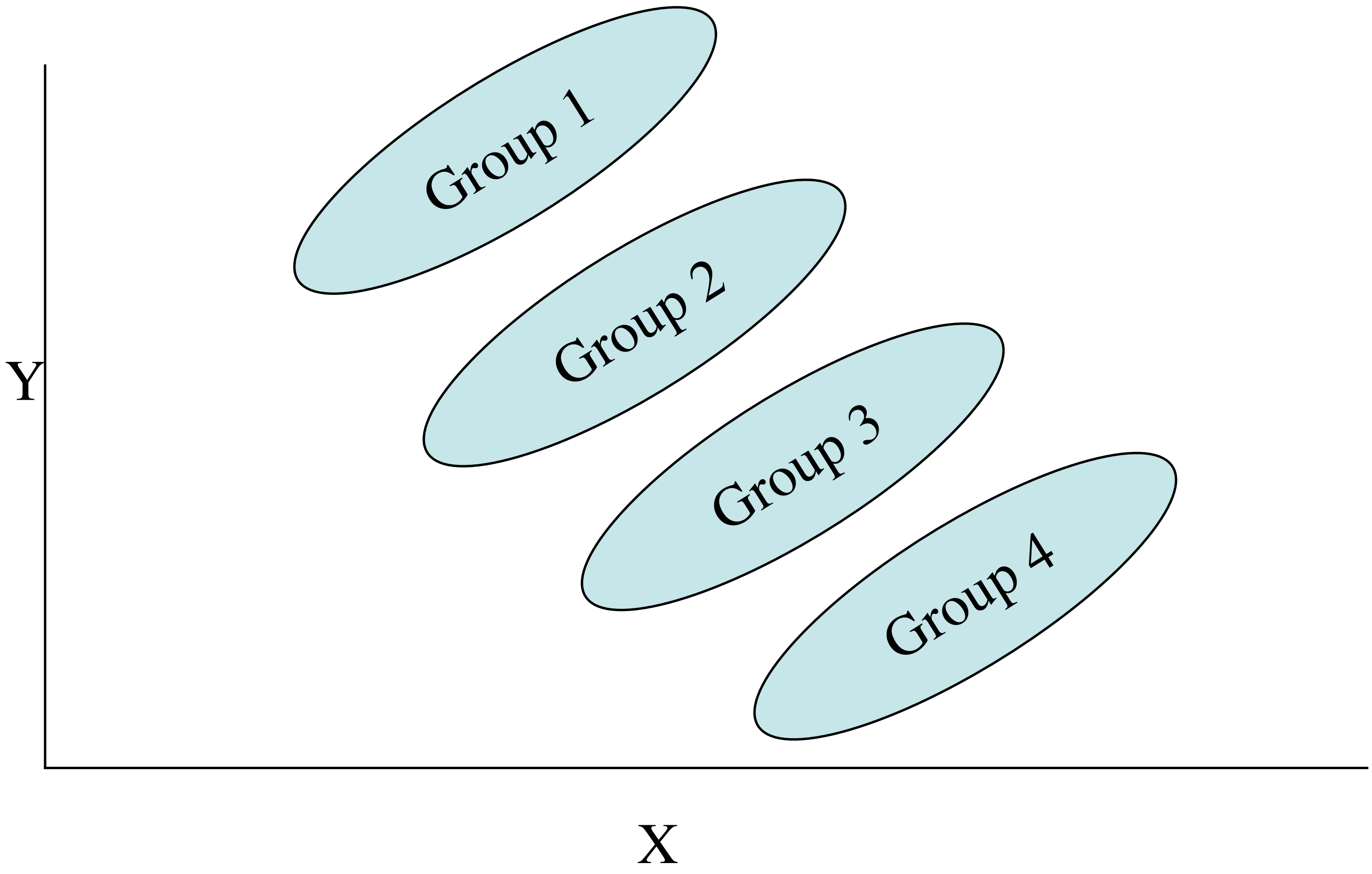
	Admit	Reject	Total
Male	40	10	50
Female	10	40	50
Total	50	50	100

Phi =(VP - HR*SR) /sqrt(HR*(1-HR)*(SR)*(1-SR)= -.60
polychoric rho = -.81

Sex discrimination?

	Department 1			Department 2		
	Admit	Reject	Total	Admit	Reject	Total
Male	40	5	45	0	5	5
Female	5	0	5	5	40	45
Total	45	5	50	5	45	50
Phi	.11			.11		
Pooled phi			-.6			

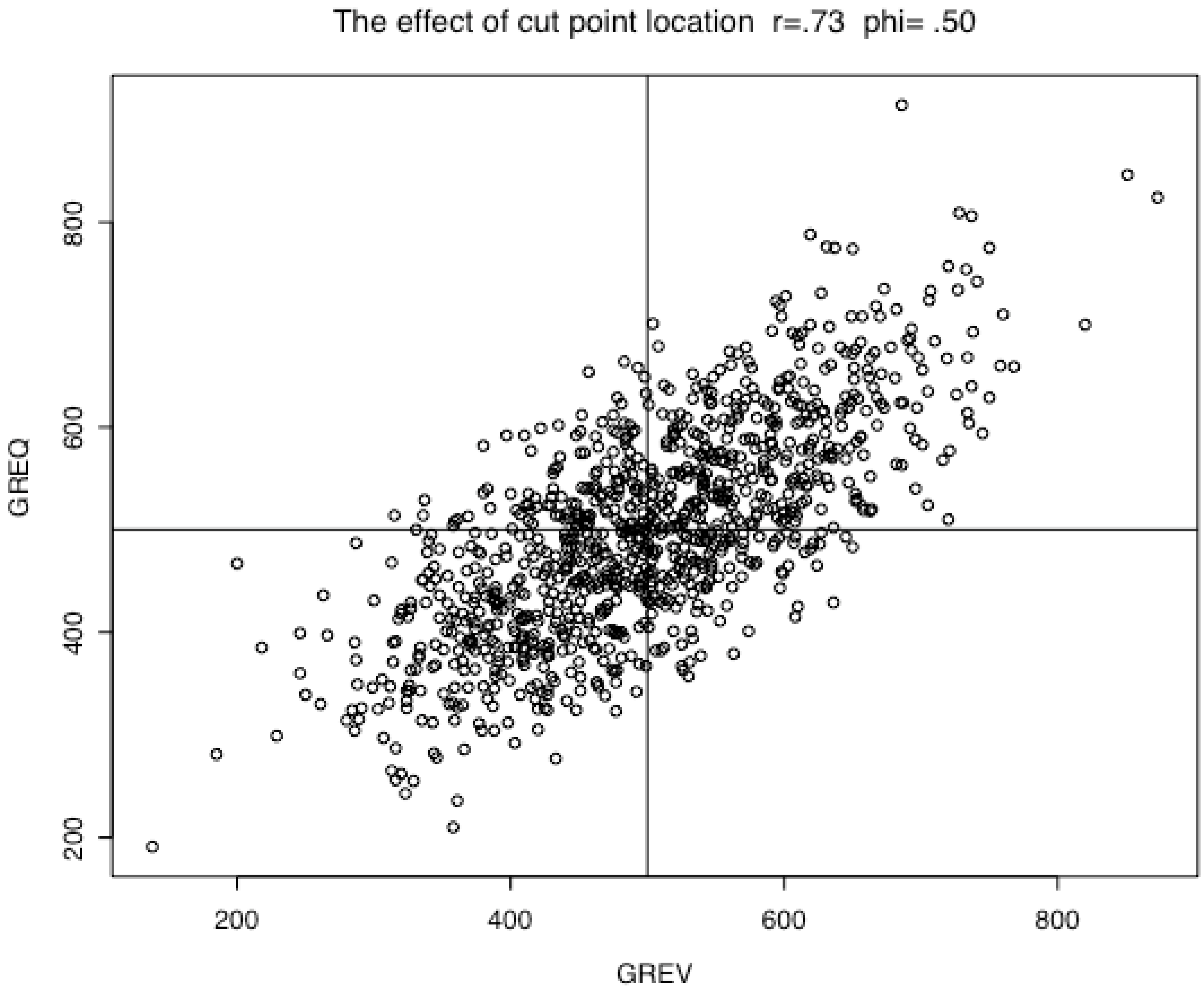
Within group vs Between Group correlation



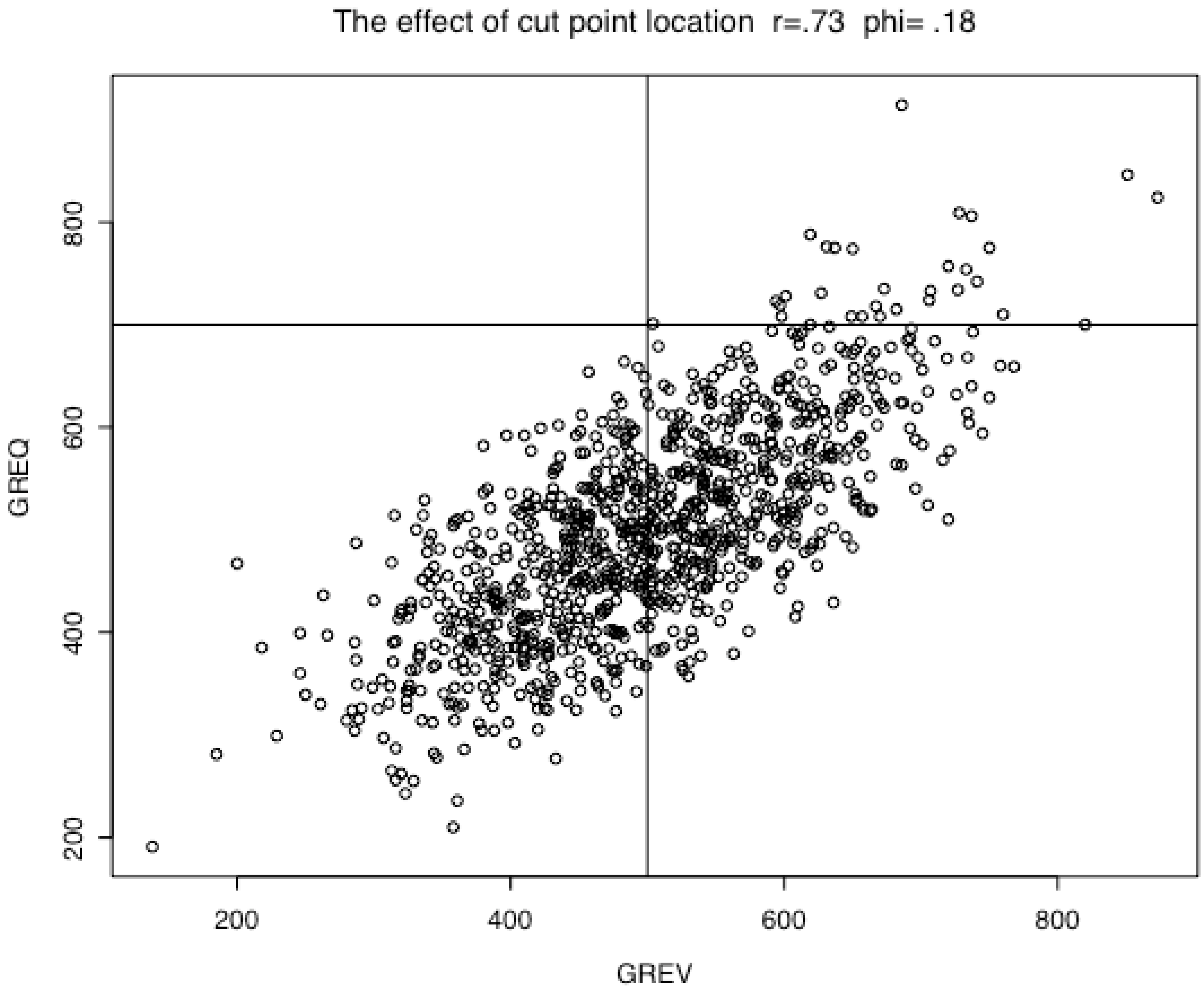
Problem Set 2

- Artificial data generated using the r-programming language
- 1000 cases with a particular structure
- First we do some simple descriptive statistics
- http://personality-project.org/revelle/syllabi/405/probset2.html

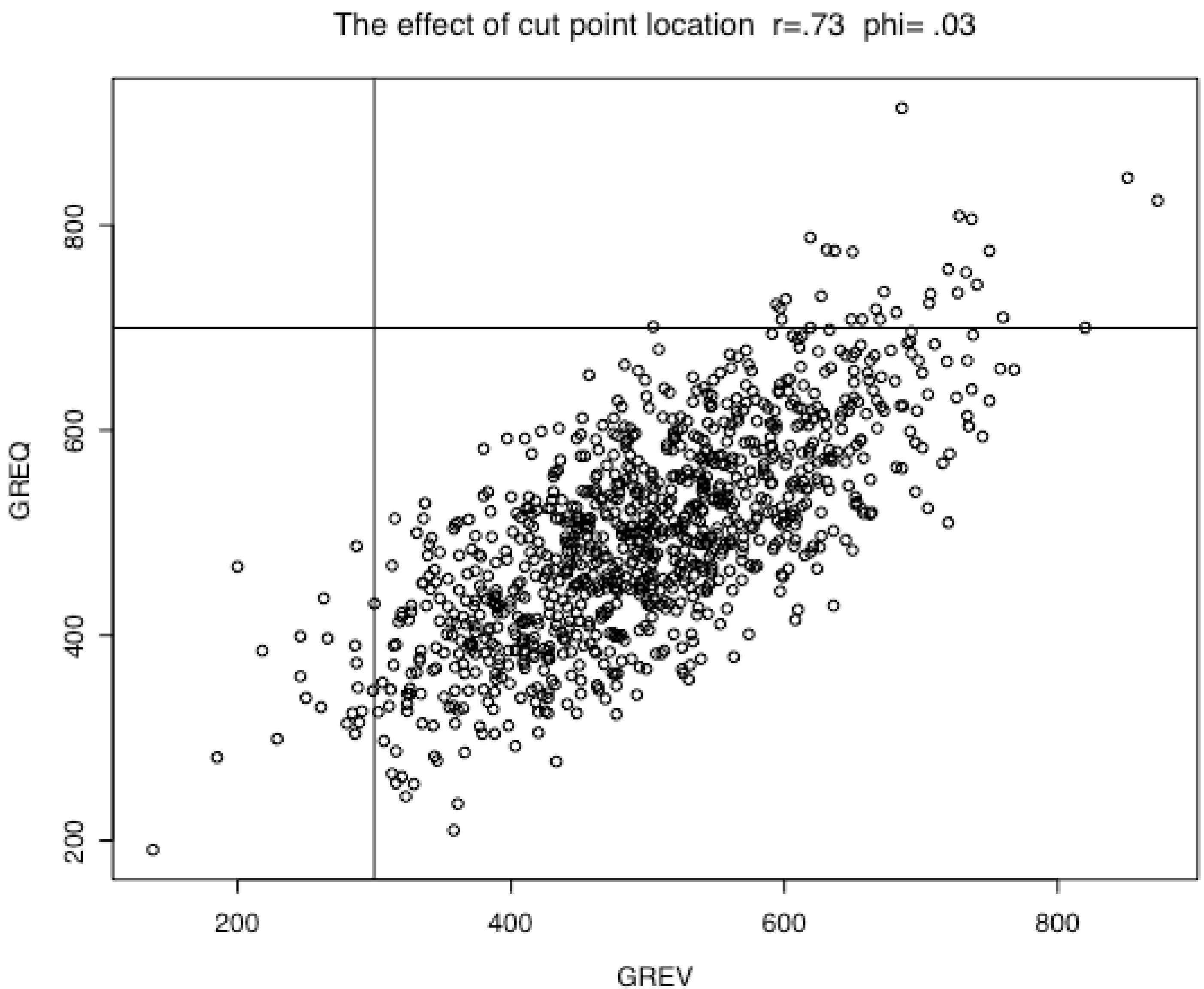
Phi vs. r the effect of cutpoints



Phi vs. r the effect of cutpoints (2)



Phi vs. r: extreme cutpoints

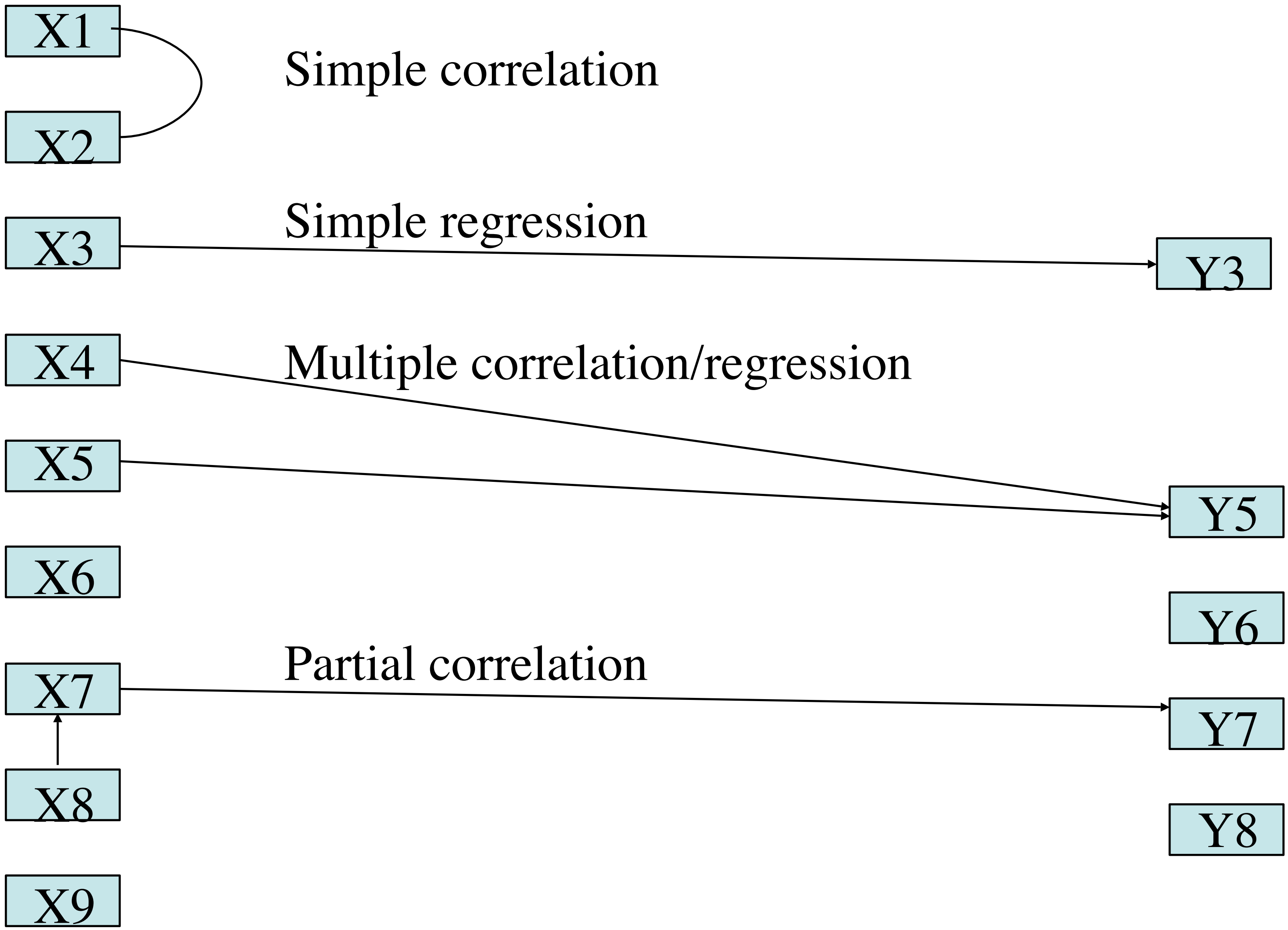


Continuous and dichotomous scales

	GREV	V2	V2l	GREQ	Q2	Q2h	GREA	GPA	MA
GREV	1.00	0.80	0.34	0.73	0.57	0.30	0.64	0.42	0.32
V2	0.80	1.00	0.15	0.58	0.50	0.18	0.51	0.37	0.23
V2l	0.34	0.15	1.00	0.21	0.15	0.03	0.19	0.15	0.12
GREQ	0.73	0.58	0.21	1.00	0.80	0.42	0.60	0.37	0.29
Q2	0.57	0.50	0.15	0.80	1.00	0.18	0.45	0.29	0.21
Q2h	0.30	0.18	0.03	0.42	0.18	1.00	0.23	0.12	0.10
GREA	0.64	0.51	0.19	0.60	0.45	0.23	1.00	0.52	0.45
GPA	0.42	0.37	0.15	0.37	0.29	0.12	0.52	1.00	0.31
MA	0.32	0.23	0.12	0.29	0.21	0.10	0.45	0.31	1.00

V2, Q2 are cut at 500
V2l is cut at 300
Q2h is cut at 700

Variance, Covariance, and Correlation

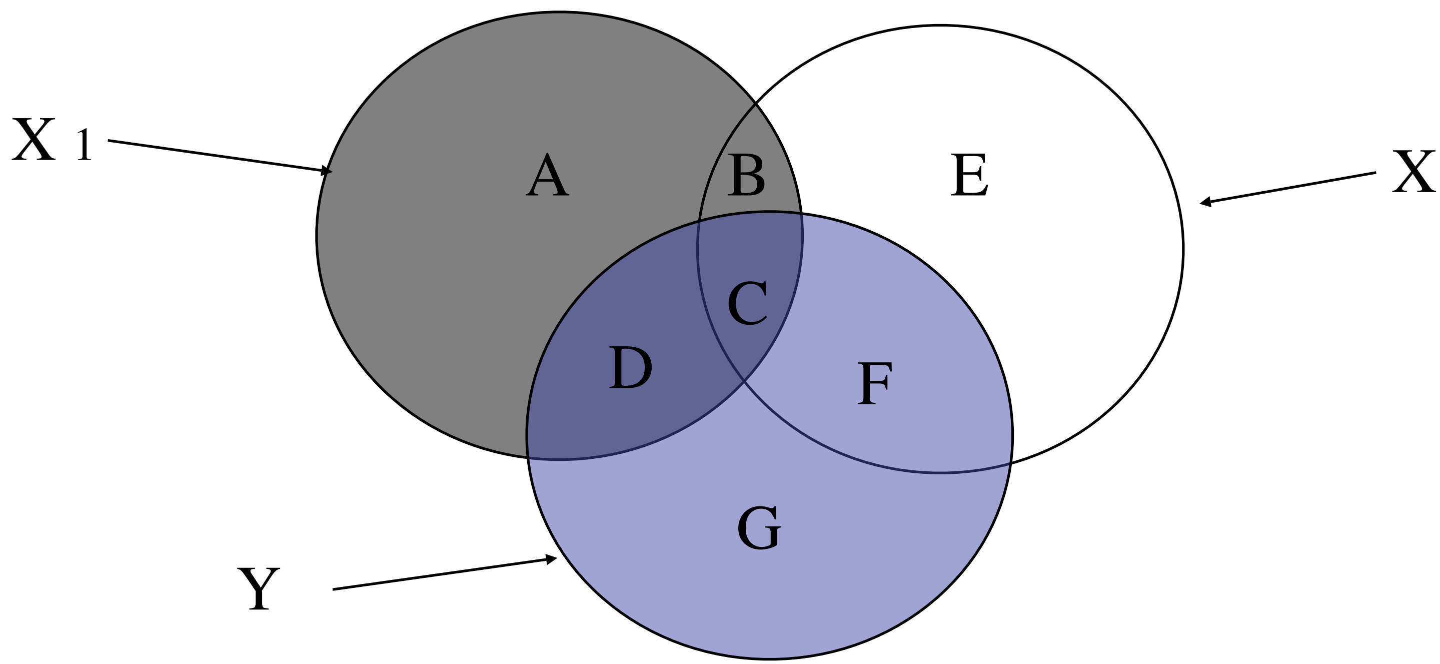


Measures of relationships with more than 2 variables

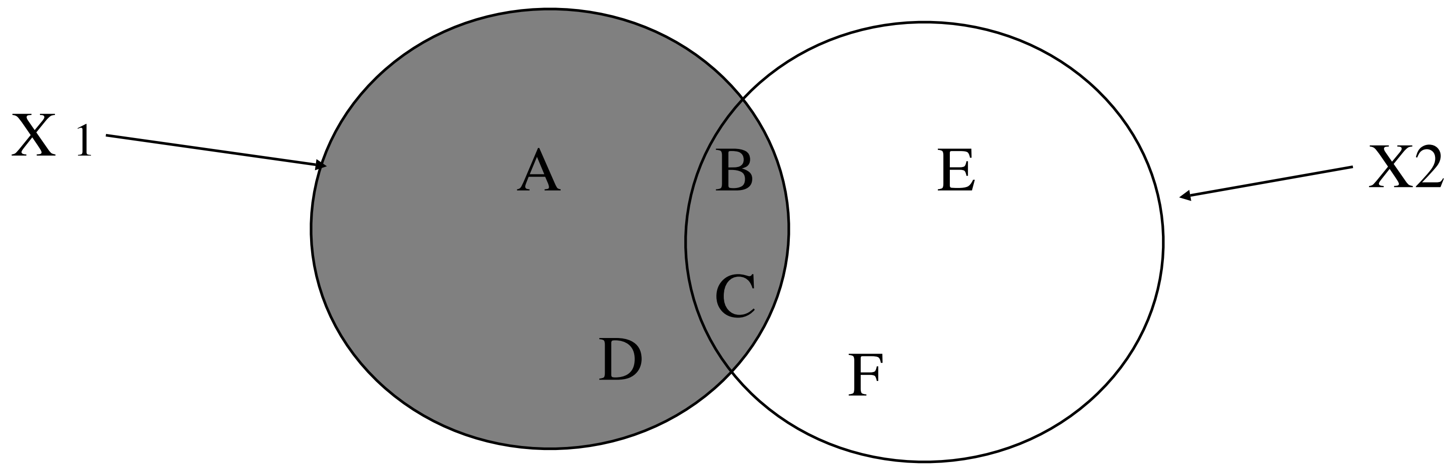
- Partial correlation
  - The relationship between x and y with z held constant (z removed)
- Multiple correlation
  - The relationship of x1 + x2 with y
  - Weight each variable by its independent contribution

Partial and Multiple Correlation

The conceptual problem



Variance, Covariance and Correlation



$V_1 = A + B + C + D$ $\quad$ $C_{12} = B + C$ $\quad$ $V_2 = E + B + C + F$

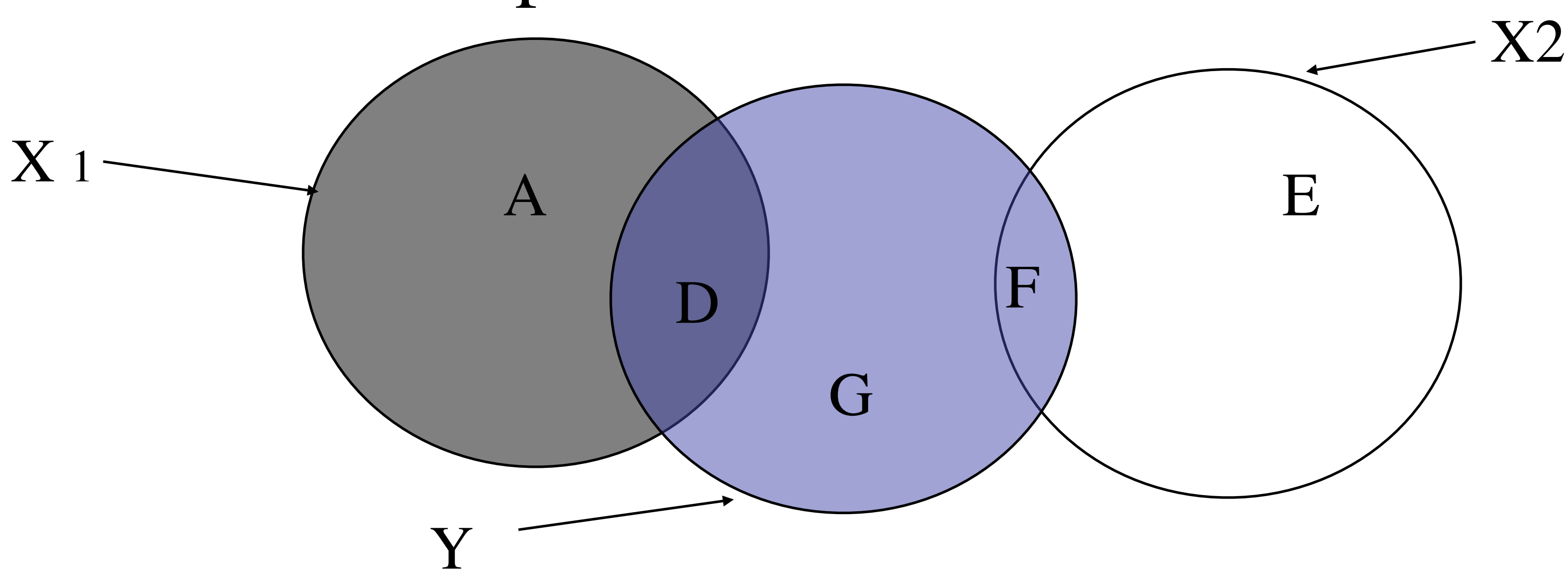
$r = C_{12}/\text{sqrt}(V_1V_2)$

$V_{1.2} = A + D$ $\quad$ $V_{2.1} = E + F$

$V_{1.2} = V_1(1\text{-}r^2)$ $\quad$ $V_{2.1} = V_2(1\text{-}r^2)$

Multiple Correlation

Independent Predictors



$V_1 = A + B + C + D$

$V_2 = E + B + C + F$

$V_Y = D + C + F + G$

$V_{1.2} = A + D$

$C_{12} = B + C$

$C_{1Y} = C + D$

$C_{2Y} = C + F$

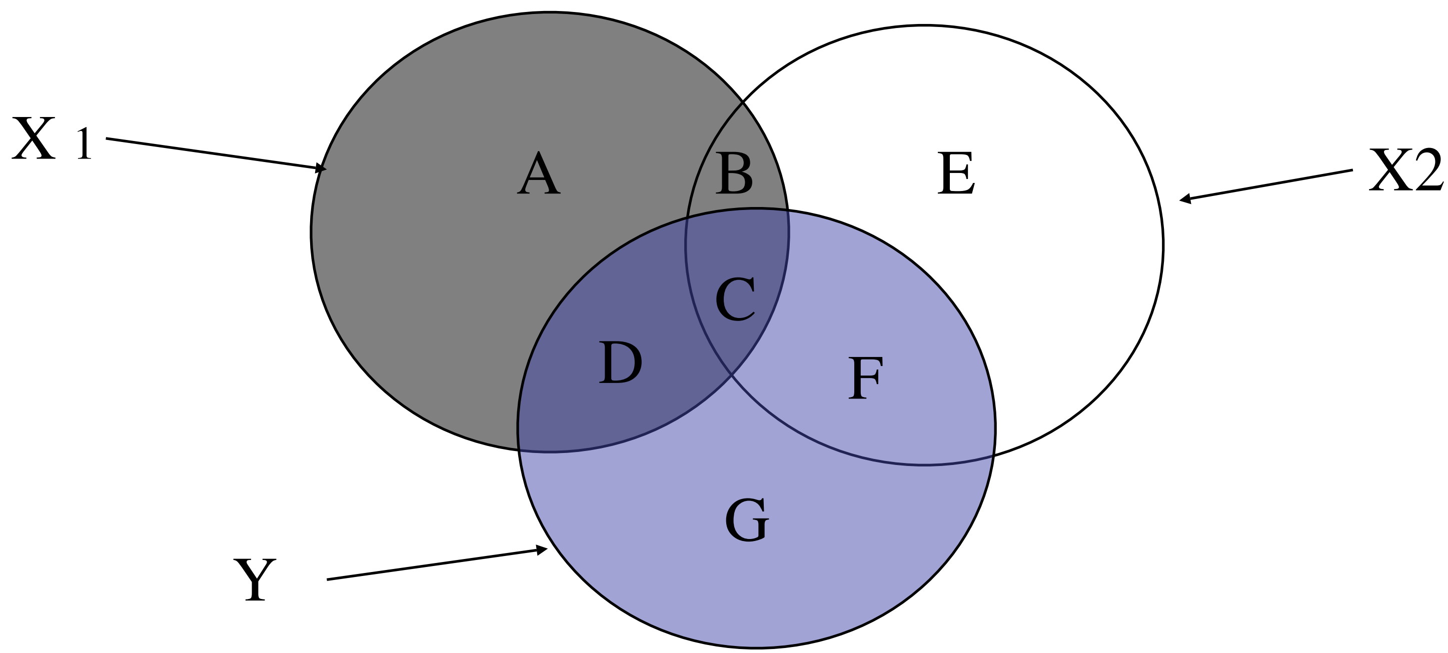
$V_{2.}1 = E + F$

$C_{1Y.2} = D$

$C_{2Y.1} = F$

$C_{(12)Y} = D + C + F$

Partial and Multiple Correlation



$V_1 = A + B + C + D$

$V_2 = E + B + C + F$

$V_Y = D + C + F + G$

$V_{1.2} = A + D$

$C_{12} = B + C$

$C_{1Y} = C + D$

$C_{2Y} = C + F$

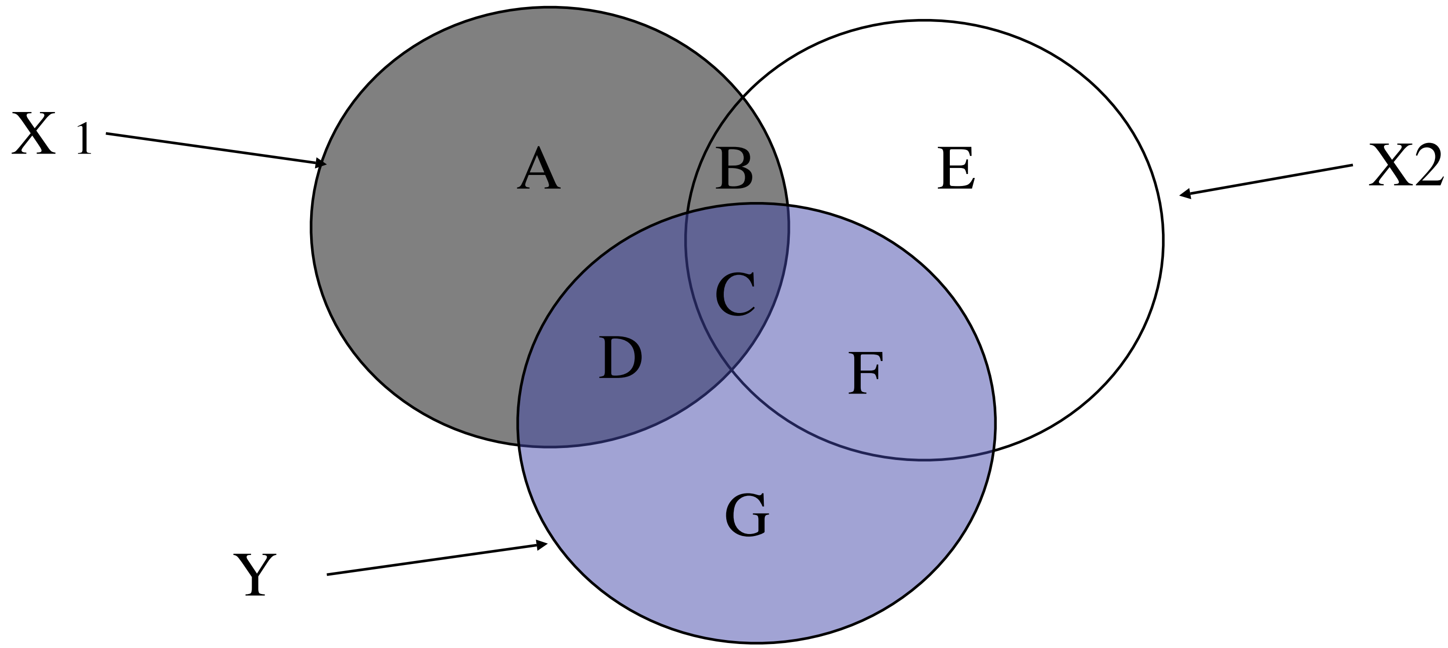
$V_{2.}1 = E + F$

$C_{1Y.2} = D$

$C_{2Y.1} = F$

$C_{(12)Y} = D + C + F$

Partial and Multiple Correlation: Partial Correlations



$V_1 = A + B + C + D$

$V_2 = E + B + C + F$

$V_Y = D + C + F + G$

$V_{1.2} = A + D$

$C_{12} = B + C$

$C_{1Y} = C + D$

$C_{2Y} = C + F$

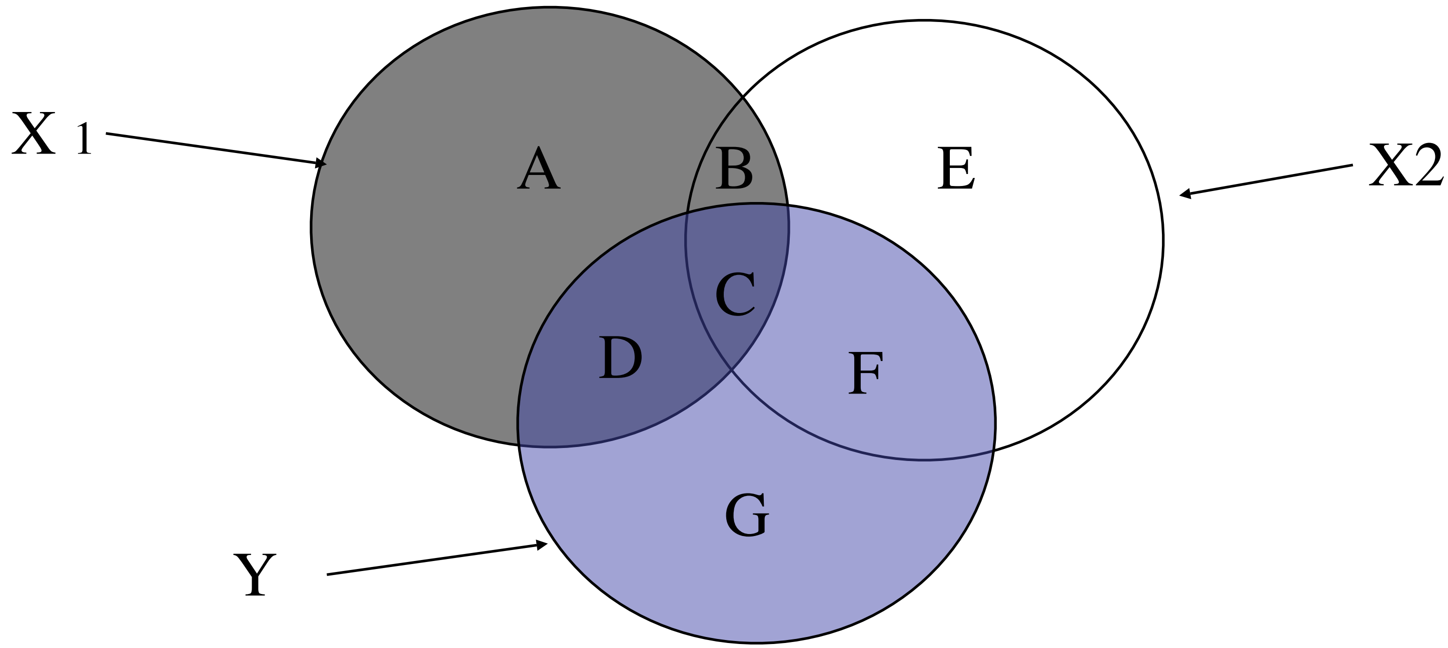
$V_{2.1} = E + F$

$C_{1Y.2} = D$

$C_{2Y.1} = F$

$$r_{1Y.2} = \frac{(r_{1y} - r_{12} * r_{2Y})}{\mathrm{sqrt}((1-r_{12}^2)*(1-r_{y2}^2))}$$

Partial and Multiple Correlation: Multiple Correlation-correlated predictors



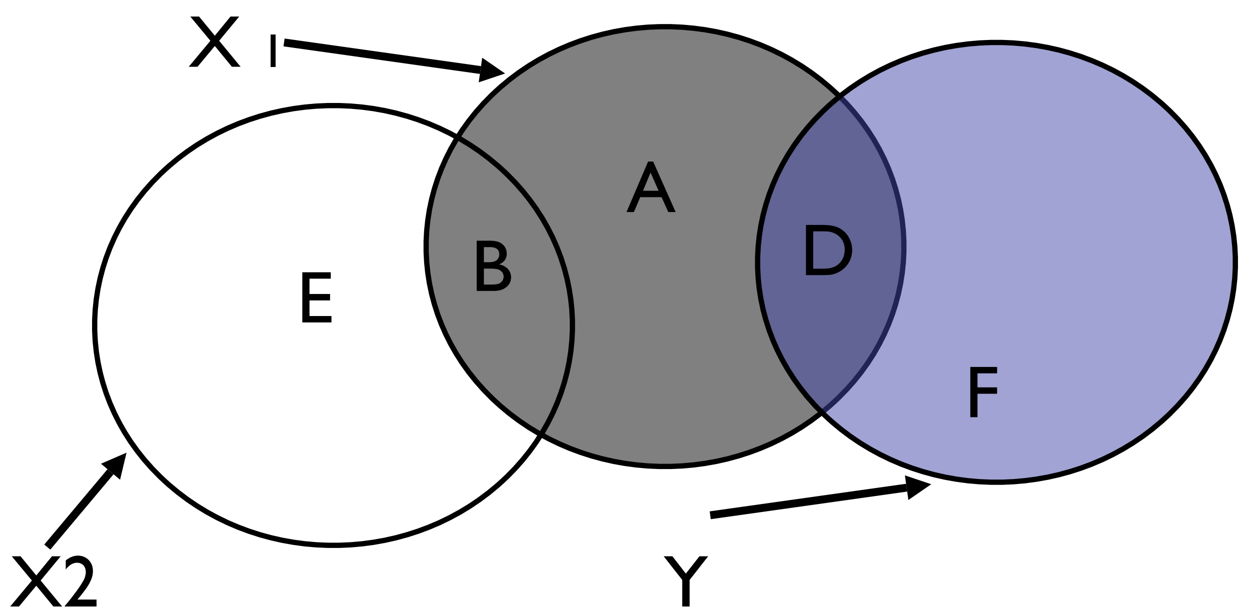
$Y = b_1X_1 + b_2X_2$

$b_1 = (r_{x1y} - r_{12}*r_{2y})/(1-r_{12}^2)$

$b_2 = (r_{x2y} - r_{12}*r_{1y})/(1-r_{12}^2)$

$R^2 = b_1r_1+b_2r_2$

Multiple Correlation:



$V_1 = A + B + C + D$

$V_2 = E + B + C + F$

$V_Y = D + C + F + G$

$V_{1.2} = A + D$

$C_{12} = B + C$

$C_{1Y} = C + D$

$C_{2Y} = C + F$

$V_{2.1} = E + F$

$C_{1Y.2} = D$

$C_{2Y.1} = F$

$C_{(12)Y} = D + C + F$

Multiple Correlation as an unweighted composite

	X_1	X_2	Y
X_1	Vx_1	Cx_1x_2	Cx_1y
X_2	Cx_1x_2	Vx_2	Cx_2y
Y	Cx_1y	Cx_2y	Vy

$Vx_1x_2 = Vx_1 + Vx_2 + 2Cx_1x_2$

$C(x_1x_2)y = Cx_1y + Cx_2y$

$$R(x_1x_2)y = \frac{C(x_1x_2)y}{Sqrt(Vx_1x_2)*V_y}$$

Multiple Correlation as a weighted composite

	b_1X_1	b_2X_2	Y
b_1X_1	$b_1^2Vx_1$	$b_1b_2Cx_1x_2$	b_1Cx_1y
b_2X_2	$b_1b_2Cx_1x_2$	$b_2^2Vx_2$	b_2Cx_2y
Y	b_1Cx_1y	b_2Cx_2y	Vy

$$R(b_1x_1b_2x_2)y = \frac{C(b_1x_1b_2x_2)}{Sqrt(Vb_1x_1b_2x_2)*V_y}$$

$Vb_1x_1b_2x_2 = b_1^2Vx_1 + b_2^2Vx_2 + 2C\ b_1b_2Cx_1x_2$

$C(b_1x_1b_2x_2)y = b_1Cx_1y + b_2Cx_2y$

Multiple Correlation as a weighted composite

	b_1X_1	b_2X_2	Y
b_1X_1	$b_1^2Vx_1$	$b_1b_2Cx_1x_2$	b_1Cx_1y
b_2X_2	$b_1b_2Cx_1x_2$	$b_2^2Vx_2$	b_2Cx_2y
Y	b_1Cx_1y	b_2Cx_2y	Vy

$$R(b_1x_1b_2x_2)y = \frac{C(b_1x_1b_2x_2)}{\text{Sqrt}(Vb_1x_1b_2x_2)*V_y}$$

Problem: Find b1, b2 to maximize R

$b_1 = (r_{x1y} - r_{12}*r_{2y})/(1-r_{12}^2)$

$b_2 = (r_{x2y} - r_{12}*r_{1y})/(1-r_{12}^2)$

Multiple regression: Matrix approach

$$
Y = X * b + e
$$

$$
\begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_i \\ \dots \\ Y_n \end{bmatrix}
=
\begin{bmatrix}
1 & x_{11} & x_{21} & \dots & x_{k1} \\
1 & x_{12} & x_{22} & \dots & x_{k2} \\
1 & \dots & \dots & \dots & \dots \\
1 & x_{1i} & x_{2i} & \dots & x_{ki} \\
1 & \dots & \dots & \dots & \dots \\
1 & x_{1n} & x_{2n} & \dots & x_{kn}
\end{bmatrix}
*
\begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \dots \\ b_k \end{bmatrix}
+
\begin{bmatrix} e_1 \\ e_2 \\ \dots \\ e_i \\ \dots \\ e_n \end{bmatrix}
$$

Matrix Algebra: a review

- Matrix algebra as a convenient notation for statistics
- Consider a matrix ${}_nX_m$ with n rows and m columns and elements x_{ij}
- Then X' (read X transpose) has m rows and n columns: ${}_mX'_n$ and elements $x_{ij}' = x_{ji}$
- ${}_mS_m = {}_mX'_n\ {}_nX_m$ is a m * m matrix of the sums (over n) of products with elements = $s_{ij} = \sum x_{ki} * x_{kj}$
- Note that if the number of columns = 1, then X is a vector with n rows. Then X'X = sum squares of x and XX' is a matrix of the products of x

Matrix Algebra: a review (2)

- The identity matrix, ${}_nI_n$ has 1's on the diagonal and 0 elsewhere.

  $IX = XI = X$

- Matrix multiplication is associative but not commutative:

  $(XY)Z = X(YZ)$ but $XY \neq YX$

- For a square matrix, X, the inverse, X^{-1} is that matrix, which when multiplied by X is I:

  $X^{-1} X = X X^{-1} = I$

Matrix Algebra: a review (3)

- Finding the inverse X^{-1} of X
- $X = IX$
- multiply both sides by a transformation with the goal of converting the left side to the Identity matrix:
  - $T_1X = T_1IX$
  - $T_2T_1X = T_2T_1IX$ until
  - $T_n \ldots T_2T_1X = I = T_n \ldots T_2T_1IX$ then
  - $(T_n \ldots T_2T_1)X = I \iff (T_n \ldots T_2T_1) = X^{-1}$

finding the inverse

$$\begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1-r^2 & 0 \\ r & 1 \end{bmatrix} = \begin{bmatrix} 1 & -r \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 \\ r & 1 \end{bmatrix} = \begin{bmatrix} 1/(1-r^2) & -r/(1-r^2) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1/(1-r^2) & -r/(1-r^2) \\ -r/(1-r^2) & 1/(1-r^2) \end{bmatrix} \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$$

Multiple regression: Matrix approach

- $Y = X * b + e$ (Y a vector, X a matrix)
- $X'Y = X'X\, b + X'e$
- $Cov_{xy} = R_{xx}b + Cov_{xe}$ (for standardized X, Y)
- Find that value of b that minimizes $||e||$
- $b = (X'X)^{-1} * X'Y$
- $b = R^{-1}X'Y$
- If X is a vector, then this is what we have already found: $b = Cov_{xy}/Var_x$
- The multivariate case is thus just a generalization of the univariate case

Multiple regression with matrix algebra (2)

$X'Y = X'X\,b + X'e$

$$\begin{bmatrix} r_{yx} \\ r_{yz} \end{bmatrix} = \begin{bmatrix} 1 & r_{xz} \\ r_{xz} & 1 \end{bmatrix} \begin{bmatrix} b_{xy.z} \\ b_{zy.x} \end{bmatrix}$$

$b = R^{-1}X'Y$

$$\begin{bmatrix} b_{xy.z} \\ b_{zy.x} \end{bmatrix} = \begin{bmatrix} 1/(1-r_{xz}^{2}) & -r_{xz}/(1-r_{xz}^{2}) \\ -r_{xz}/(1-r_{xz}^{2}) & 1/(1-r_{xz}^{2}) \end{bmatrix} \begin{bmatrix} r_{yx} \\ r_{yz} \end{bmatrix}$$

$$\begin{bmatrix} b_{xy.z} \\ b_{zy.x} \end{bmatrix} = \begin{bmatrix} (r_{yx}-r_{yz}\,r_{xz})/(1-r_{xz}^{2}) \\ (r_{yz}-r_{yx}\,r_{xz})/(1-r_{xz}^{2}) \end{bmatrix}$$

Compare this to the solution derived earlier

Correlation and Regression as path models or matrix models

I. Path notation shows Pattern of relationships

   A. path arithmetic

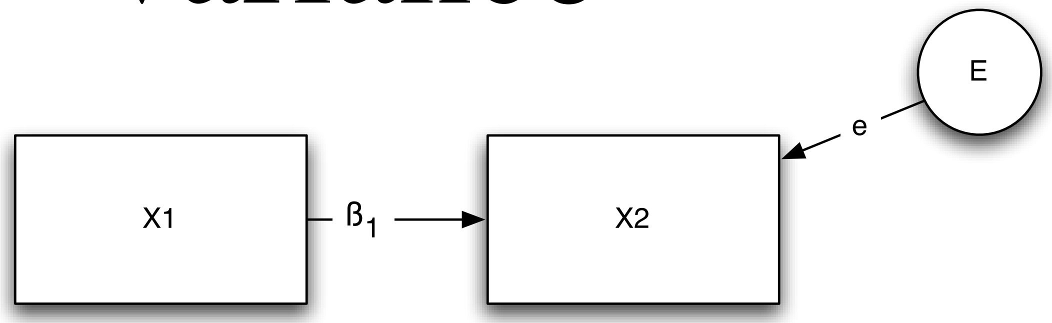
      1. no loops

      2. one curved arrow/path

      3. no forward and then back

II. Matrix notation of paths can show Pattern, Structure, and represent data (and allow for calculation)

Regression: Modeling the variance



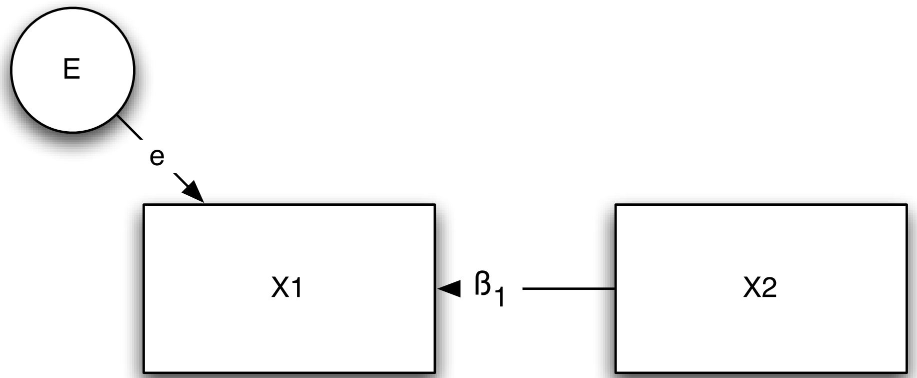
Pattern

	X1	E
X1	1	0
X2	ß	e

PP'

	X1	X2
X1	1	ß
X2	ß	$ß^2+e^2$

Correlation or regression: which way is the direction



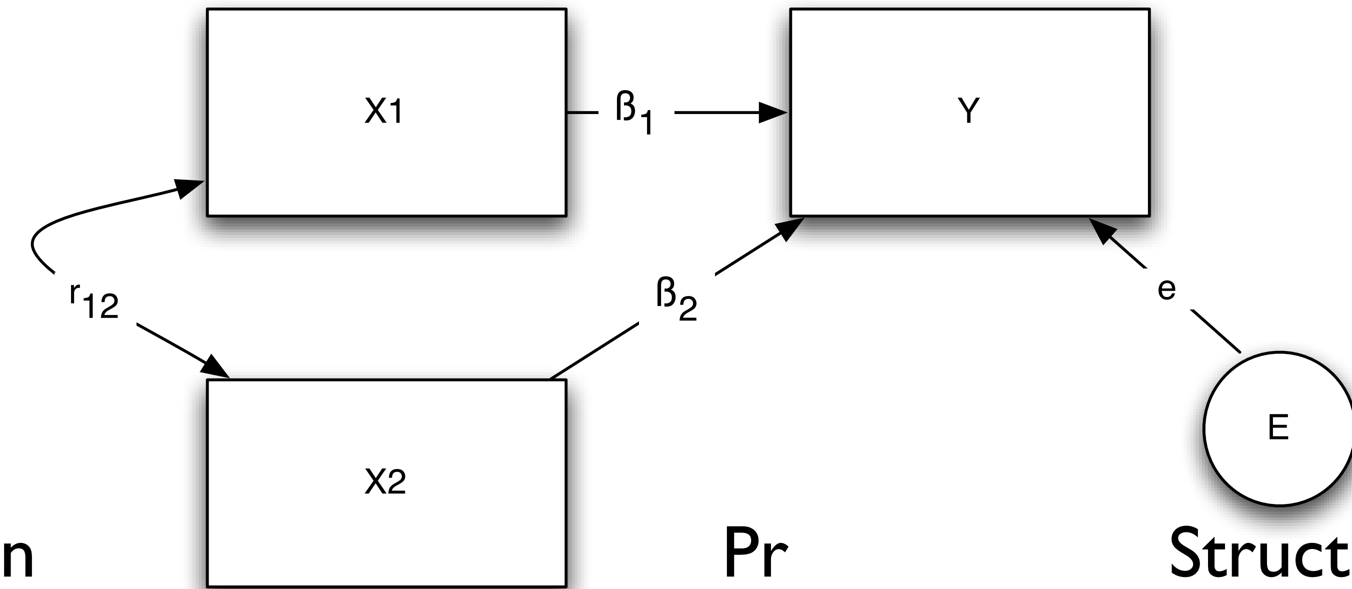
Pattern

	X2	E
X1	ß	0
X2	1	e

PP'

	X1	X2
X1	$ß^2+e^2$	ß
X2	ß	1

Multiple regression



Pattern

	X_1	X_2	E
X_1	1	0	0
X_2	0	1	0
Y	$ß_1$	$ß_2$	e

Pr

	X_1	X_2	E
X_1	1	r_{12}	0
X_2	r_{12}	1	0
E	0	0	1

Structure

	X_1	X_2	Y
X_1	1	r_{12}	0
X_2	r_{12}	1	0
Y	$ß_1+ß_2r$	$ß_1r+ß_2$	e

Multiple Regression as a set of simultaneous equations

$$\left\{ \begin{array}{ccc} r_{x1x1} & r_{x1x2} & r_{x1y} \\ r_{x1x2} & r_{x2x2} & r_{x2y} \\ r_{x1y} & r_{x2y} & r_{yy} \end{array} \right\}$$

$$\left\{ \begin{array}{c} r_{x1x1}\beta_1 + r_{x1x2}\beta_2 = r_{x1y} \\ r_{x1x2}\beta_1 + r_{x2x2}\beta_2 = r_{x2y} \end{array} \right\}.$$

$$\left\{ \begin{array}{c} \beta_1 = (r_{x1y}r_{x2x2} - r_{x1x2}r_{x2y})/(r_{x1x1}r_{x2x2} - r_{x1x2}^2) \\ \beta_2 = (r_{x2y}r_{x1x1} - r_{x1x2}r_{x1y})/(r_{x1x1}r_{x2x2} - r_{x1x2}^2) \end{array} \right\}$$

Matrix representation

$$(\beta_1 \beta_2) \begin{pmatrix} r_{x1x1} & r_{x1x2} \\ r_{x1x2} & r_{x2x2} \end{pmatrix} = (r_{x1y} \quad r_{x2x2})$$

$$\beta = (\beta_1 \beta_2),\ \mathrm{R} = \begin{pmatrix} r_{x1x1} & r_{x1x2} \\ r_{x1x2} & r_{x2x2} \end{pmatrix} \text{ and } r_{xy} = (r_{x1y} \quad r_{x2x2})$$

$$\beta R = r_{xy}$$

$$\beta = \beta R R^{-1} = r_{xy} R^{-1}$$

Finding the inverse

$$R = IR$$

$$\begin{pmatrix} r_{x1x1} & r_{x1x2} \\ r_{x1x2} & r_{x2x2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} r_{x1x1} & r_{x1x2} \\ r_{x1x2} & r_{x2x2} \end{pmatrix}$$

$$T_1 = \begin{pmatrix} \frac{1}{r_{11}} & 0 \\ 0 & \frac{1}{r_{22}} \end{pmatrix}$$

The inverse of a matrix

$$T_1 R = T_1 I R$$

$$\begin{pmatrix} 1 & \frac{r_{12}}{r_{11}} \\ \frac{r_{12}}{r_{22}} & 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{r_{11}} & 0 \\ 0 & \frac{1}{r_{22}} \end{pmatrix} \begin{pmatrix} r_{11} & r_{12} \\ r_{12} & r_{22} \end{pmatrix}$$

...

$$T_3 T_2 T_1 R = I = R^{-1} R$$

$$T_3 T_2 T_1 I = R^{-1}$$

The inverse of a 2 x 2

```
> R2
      x1   x2
x1 1.00 0.56
x2 0.56 1.00
> round(solve(R2),2)
      x1    x2
x1  1.46 -0.82
x2 -0.82  1.46
```

The inverse of a 3 x 3

```
> R
     x1   x2   x3
x1 1.00 0.56 0.48
x2 0.56 1.00 0.42
x3 0.48 0.42 1.00
> round(solve(R),2)
      x1    x2    x3
x1  1.63 -0.71 -0.48
x2 -0.71  1.52 -0.30
x3 -0.48 -0.30  1.36
```

Unit weights versus optimal weights - "It don't make no nevermind"

r_{x1x2}	r_{x1y}	r_{x2y}	beta 1	beta 2	R	R^2	Unit Wt	UW^2
0.0	0.5	0.5	0.50	0.50	0.71	0.50	0.71	0.50
0.3	0.5	0.5	0.38	0.38	0.62	0.38	0.62	0.38
0.5	0.5	0.5	0.33	0.33	0.58	0.33	0.58	0.33
0.7	0.5	0.5	0.29	0.29	0.54	0.29	0.54	0.29
0.3	0.5	0	0.55	-0.16	0.52	0.27	0.31	0.10
0.3	0.5	0.3	0.45	0.16	0.52	0.27	0.50	0.25

If X_1 and X_2 are both positively correlated with Y, then the effect of unit weighting versus optimal (beta) weighting is negligible. But, if one variable is not very good or zero,then unit weighting will not be as effective.

Multiple regresssion

I. At data level

A. $Y = X\beta + \partial$

B. $\beta = (X'X)^{-1} X'Y$

II. At structure level

A. $\beta = R^{-1}r_{xy}$

Multiple Regression:

$y = xb \Rightarrow b_{xy} = R^{-1} r_{xy}$

R

	x1	x2	x3
x1	1	0	0
x2	0	1	0
x3	0	0	1

r_{xy}

	y
x1	0.8
x2	0.7
x3	0.6

R^{-1}

	x1	x2	x3
x1	1	0	0
x2	0	1	0
x3	0	0	1

$R^{-1} r_{xy}$

	y
x1	0.8
x2	0.7
x3	0.6

Multiple Regression:

$$y = xb \;\; \text{->} \; b_{xy} = R^{-1} r_{xy}$$

R

	x1	x2	x3
x1	1.00	0.56	0.48
x2	0.56	1.00	0.42
x3	0.48	0.42	1.00

r_{xy}

	y
x1	0.8
x2	0.7
x3	0.6

R^{-1}

	x1	x2	x3
x1	1.63	-0.71	-0.48
x2	-0.71	1.52	-0.30
x3	-0.48	-0.30	1.36

$R^{-1} r_{xy}$

	y
x1	0.52
x2	0.32
x3	0.22

Multiple Regression: $y = xb \rightarrow b_{xy} = R^{-1}r_{xy}$

R

	x1	x2	x3
x1	1.0	0.8	0.8
x2	0.8	1.0	0.8
x3	0.8	0.8	1.0

r_{xy}

	y
x1	0.8
x2	0.7
x3	0.6

R^{-1}

	x1	x2	x3
x1	3.46	-1.54	-1.54
x2	-1.54	3.46	-1.54
x3	-1.54	-1.54	3.46

$R^{-1} r_{xy}$

	y
x1	0.77
x2	0.27
x3	-0.23

Solution space is relatively flat as f(beta)

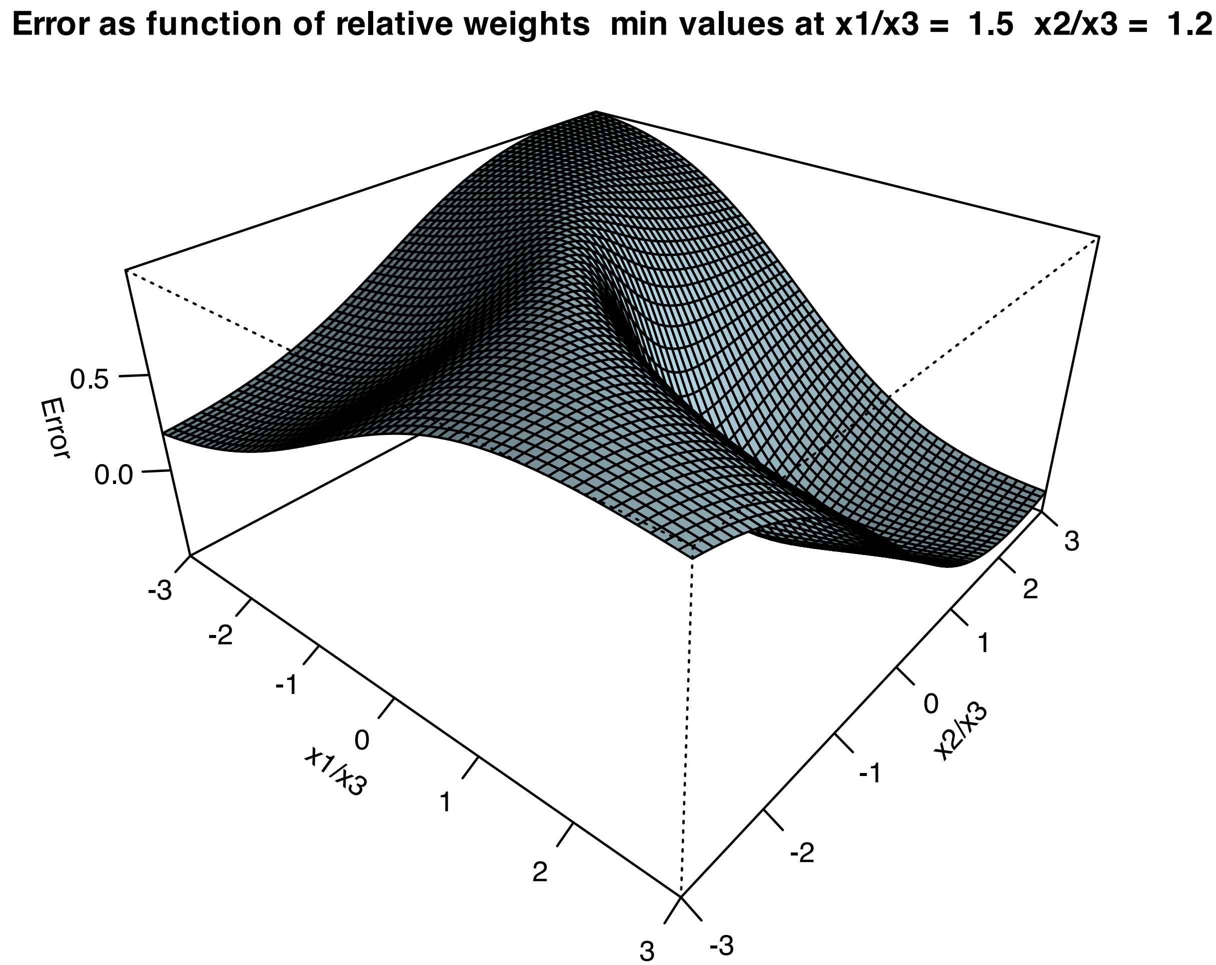
I. Although the optimal beta weights may be found precisely by multiple regression, the solution space is relatively flat and many alternative solutions are almost as good.

II. Iterative solutions can discover local minima that are far from the optimal solution

Multiple regression

	x1	x2	x3
x1	1	0	0
x2	0	1	0
x3	0	0	1

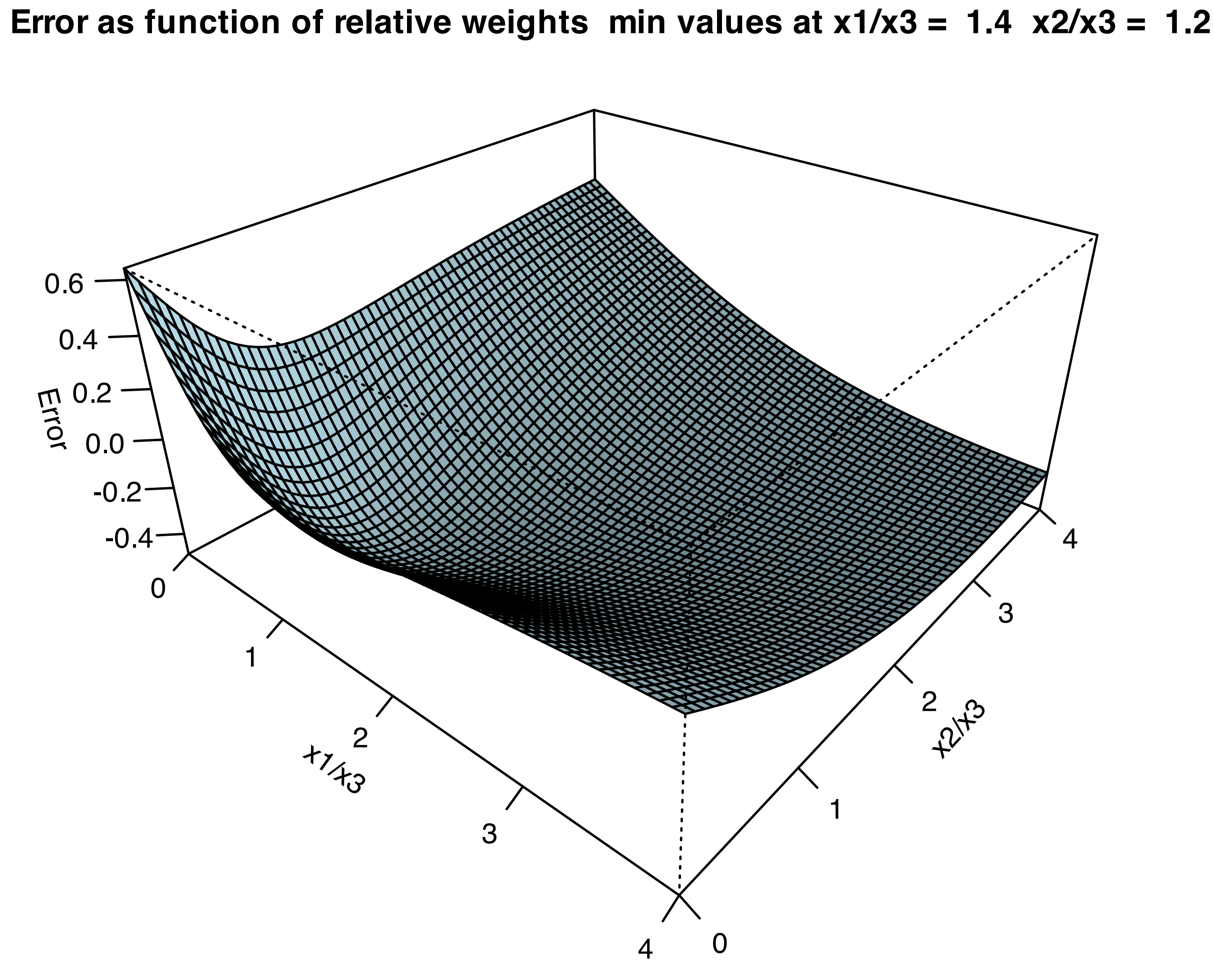
	y
x1	0.8
x2	0.7
x3	0.6



Multiple regression

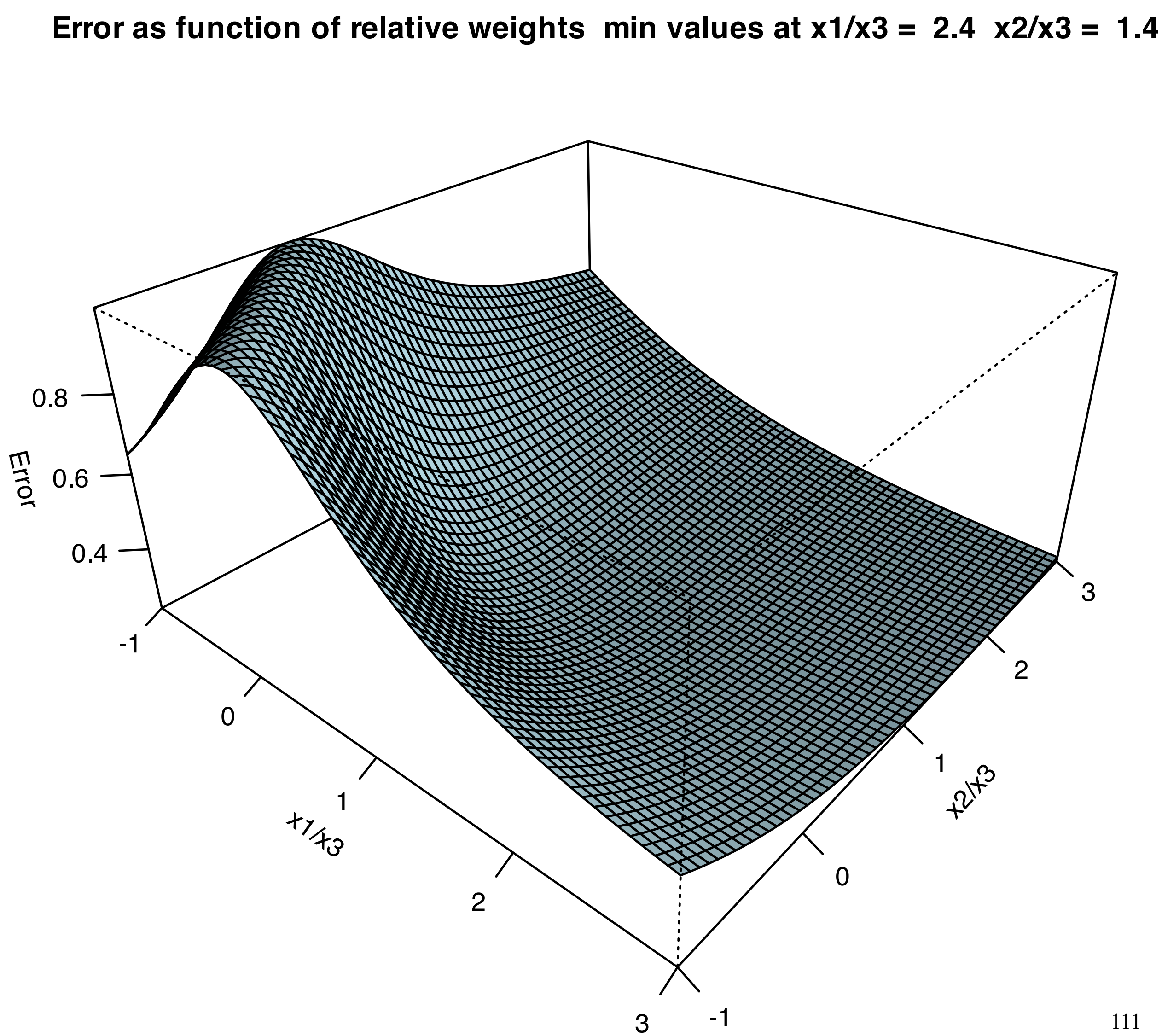
	x1	x2	x3
x1	1	0	0
x2	0	1	0
x3	0	0	1

	y
x1	0.8
x2	0.7
x3	0.6



Multiple regression

	x1	x2	x3
x1	1.00	0.56	0.48
x2	0.56	1.00	0.42
x3	0.48	0.42	1.00



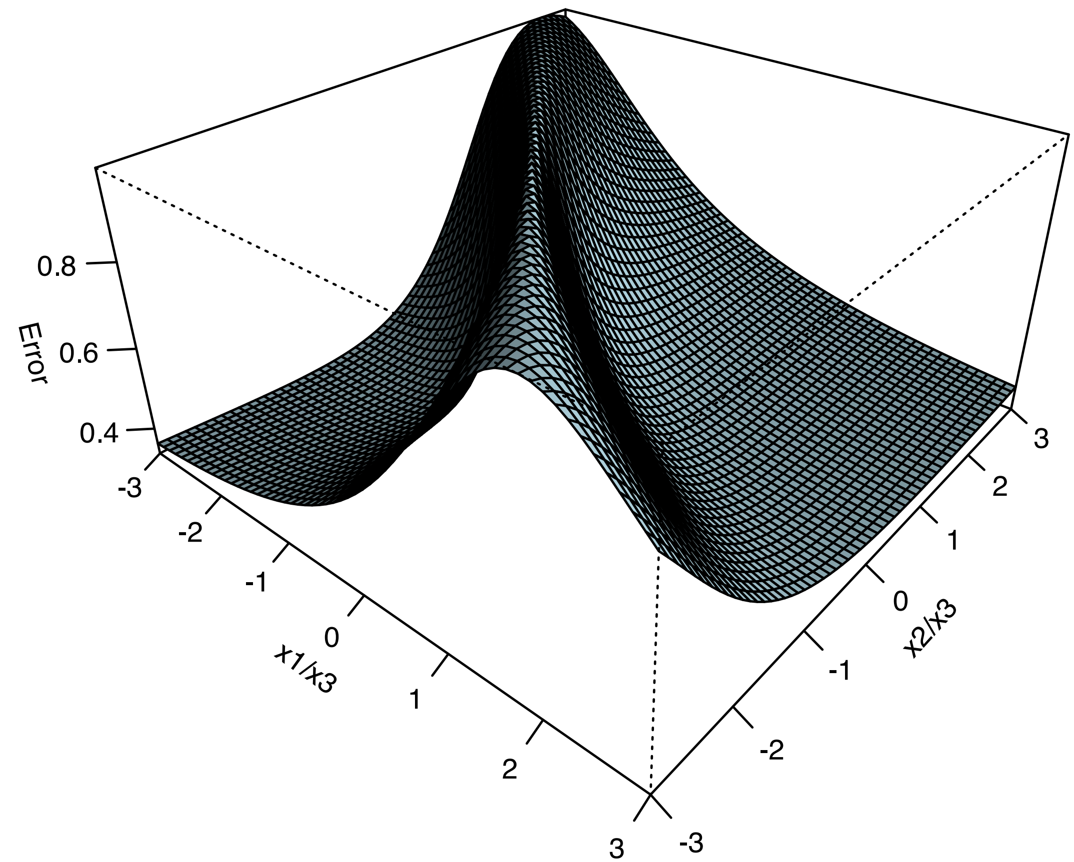
	y
x1	0.52
x2	0.32
x3	0.22

Multiple regression

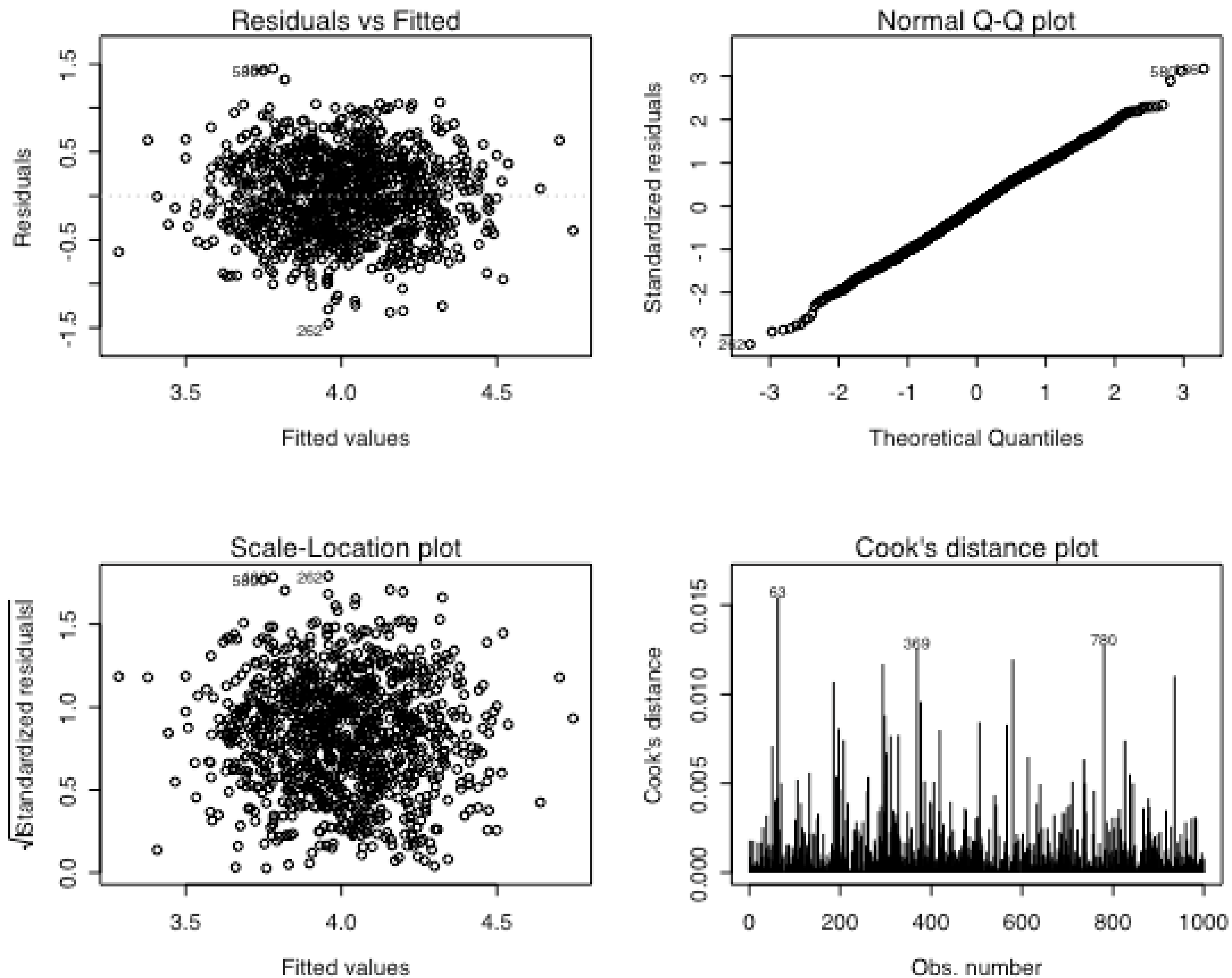
	x1	x2	x3
x1	1.0	0.8	0.8
x2	0.8	1.0	0.8
x3	0.8	0.8	1.0

	y
x1	0.77
x2	0.27
x3	-0.23

Error as function of relative weights min values at x1/x3 = -2.9 x2/x3 = -1.1



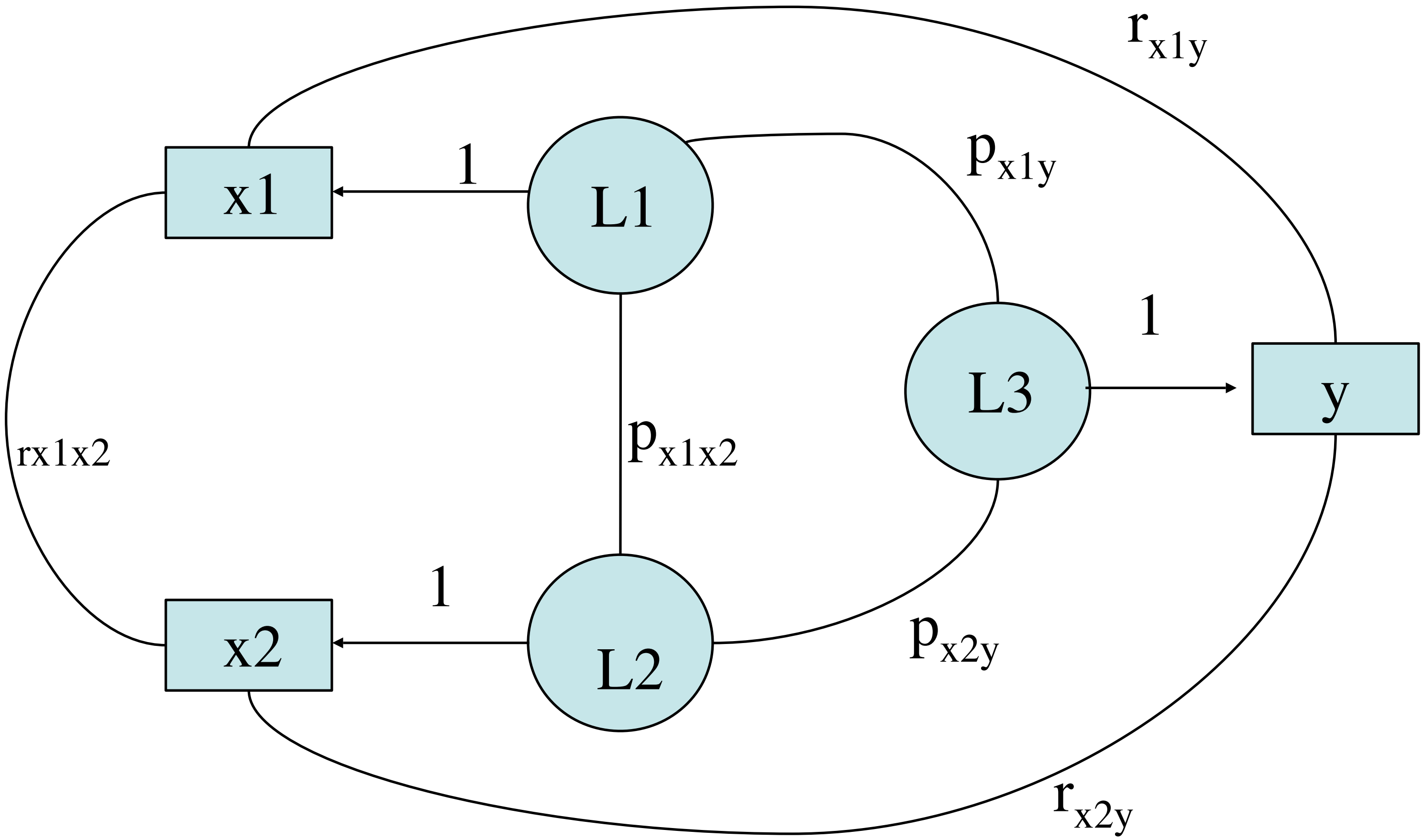
Regression diagnostics



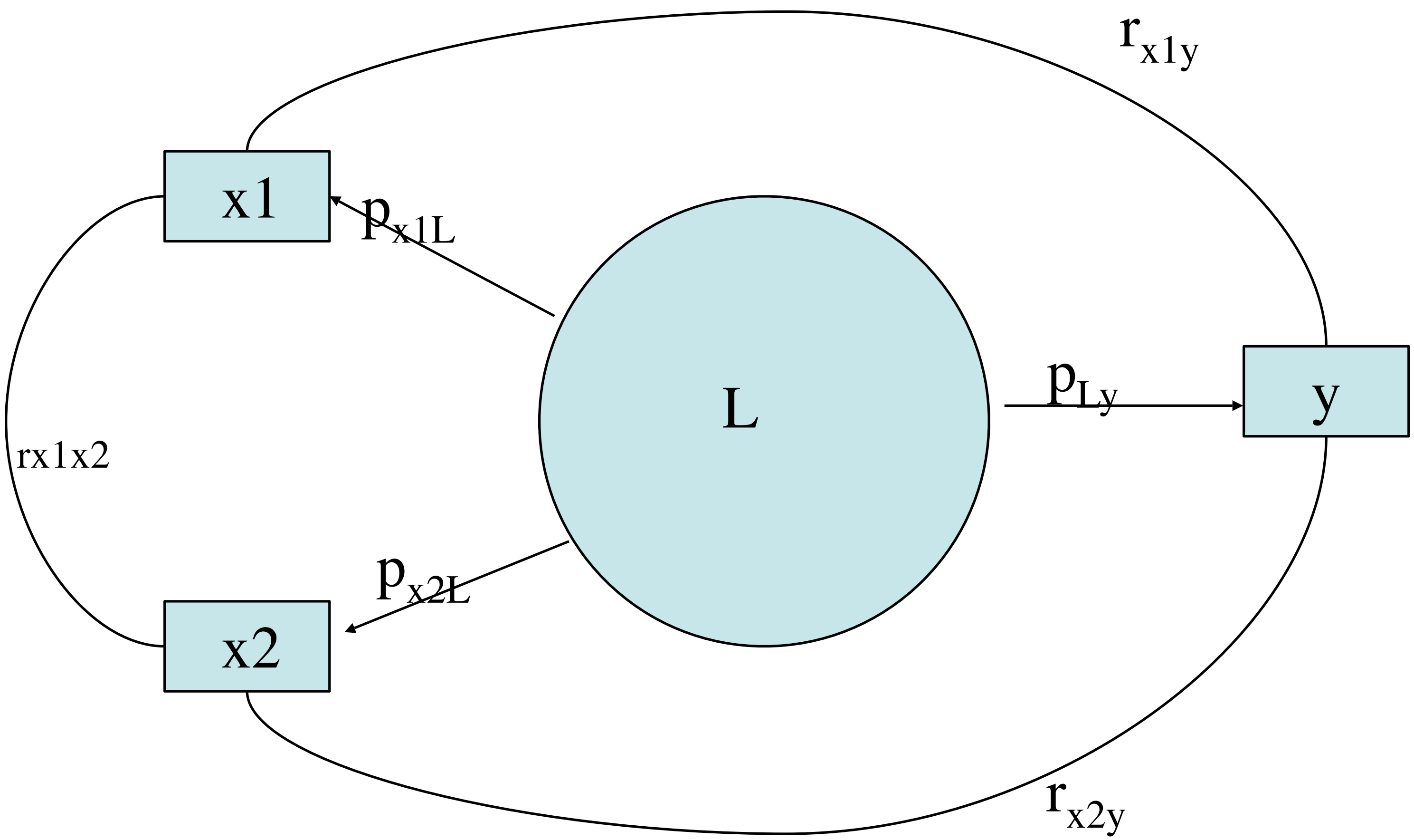
Problems with correlations

- Simpson's paradox and the problem of aggregating groups
  - Within group relationships are not the same as between group or pooled relationships
- Phi coefficients and the problem of unequal marginal distributions
- Alternative interpretations of partial correlations

Partial correlation: conventional model



Partial correlation: Alternative model



Partial Correlation: classical model

	X_1	X_2	Y
X_1	1.00		
X_2	.72	1.00	
Y	.63	.56	1.00

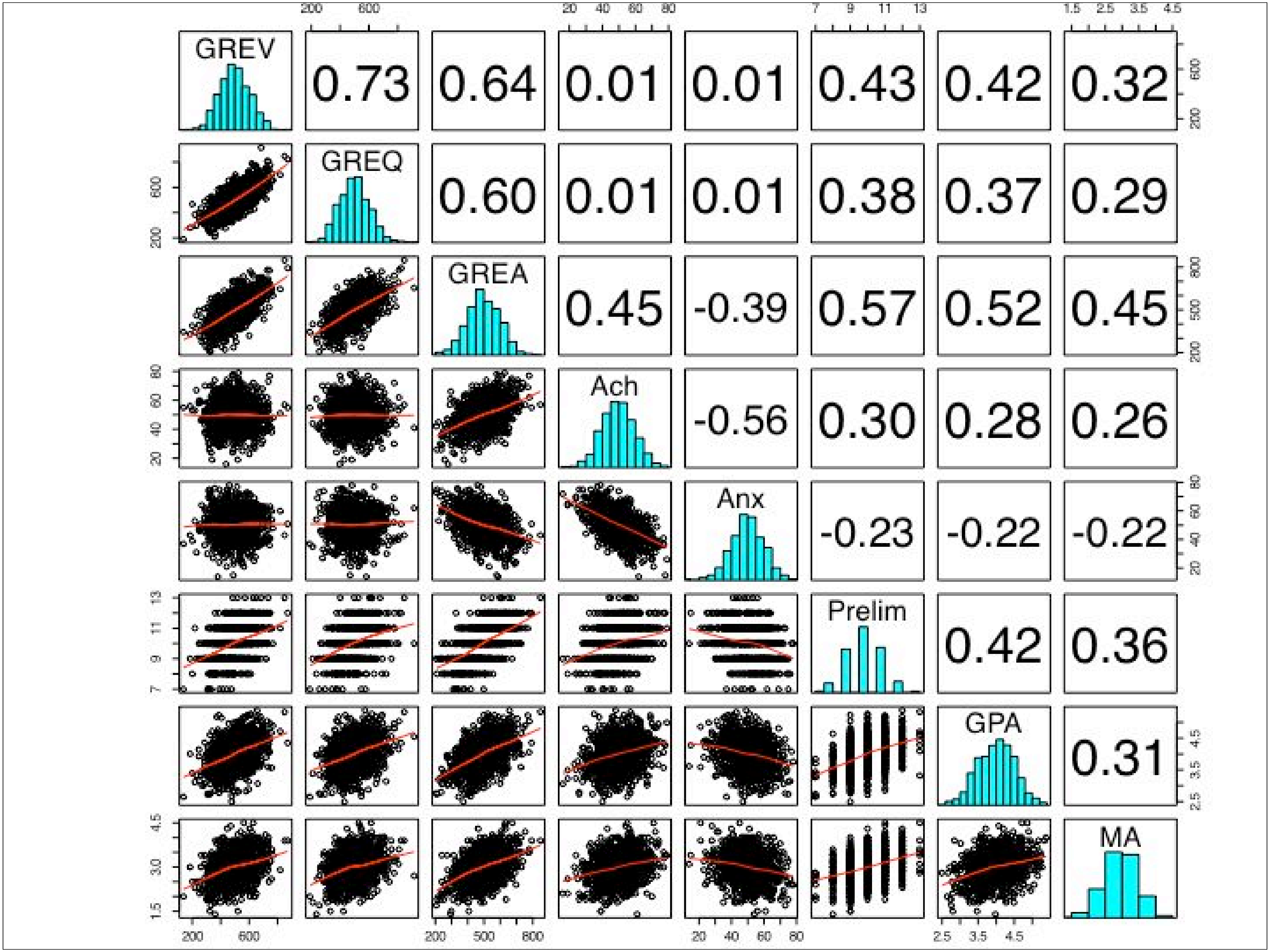
Partial r = $(r_{x1y}-r_{x1x2}*r_{x2y})/sqrt((1-r_{x1x2}{}^2)*(1-r_{x2y}{}^2))$

Rx1y.x2 = .33 (traditional model) but = 0 with structural model

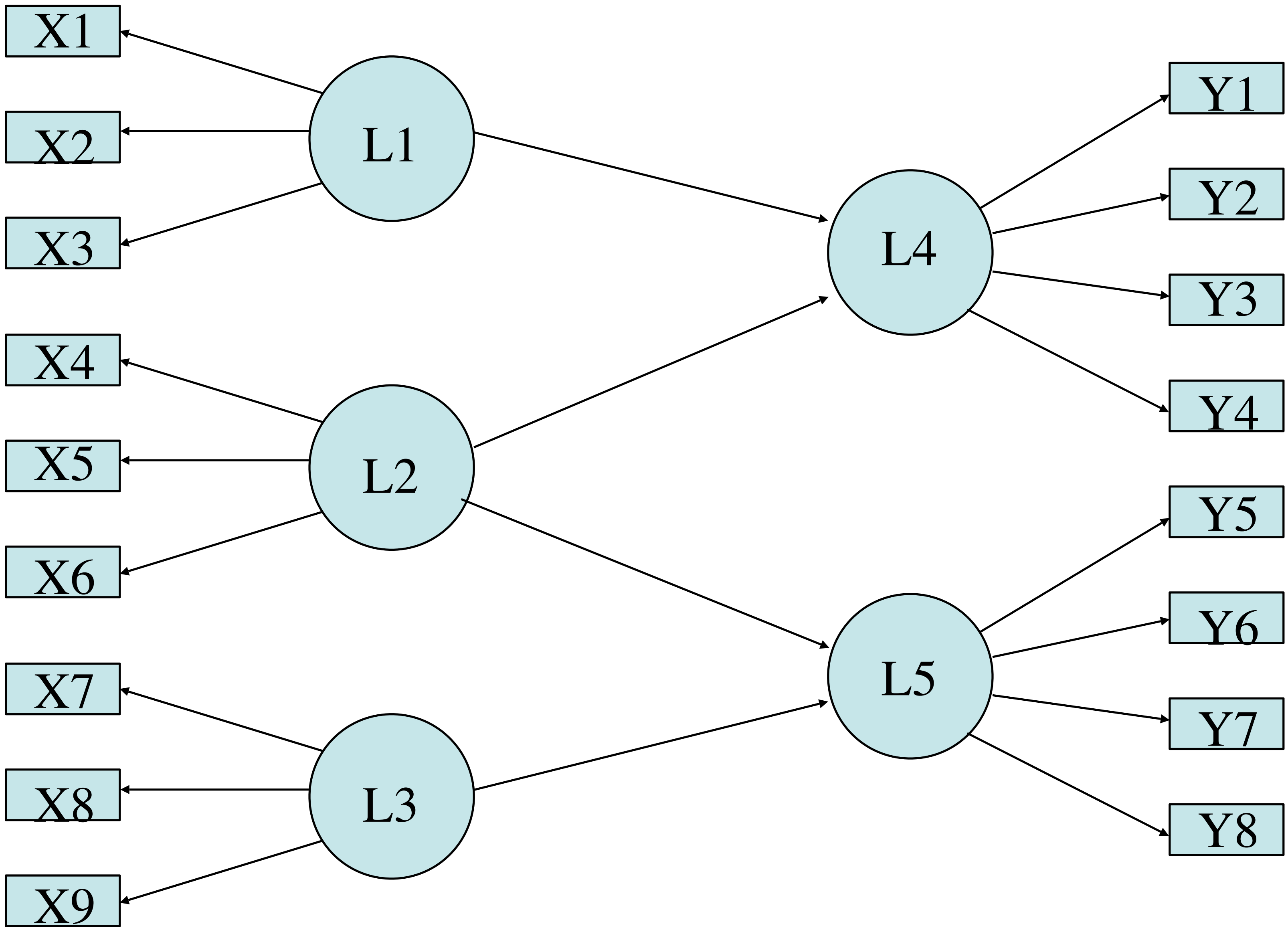
Find the correlations

```
round(cor(dataset),2)          #find the correlation matrix
        #round off to 2 decimals
```

	GREV	GREQ	GREA	Ach	Anx	Prelim	GPA	MA
GREV	1.00	0.73	0.64	0.01	0.01	0.43	0.42	0.32
GREQ	0.73	1.00	0.60	0.01	0.01	0.38	0.37	0.29
GREA	0.64	0.60	1.00	0.45	-0.39	0.57	0.52	0.45
Ach	0.01	0.01	0.45	1.00	-0.56	0.30	0.28	0.26
Anx	0.01	0.01	-0.39	-0.56	1.00	-0.23	-0.22	-0.22
Prelim	0.43	0.38	0.57	0.30	-0.23	1.00	0.42	0.36
GPA	0.42	0.37	0.52	0.28	-0.22	0.42	1.00	0.31
MA	0.32	0.29	0.45	0.26	-0.22	0.36	0.31	1.00


GREV
GREQ
GREA
Ach
Anx
Prelim
GPA
MA
0.73
0.64
0.01
0.01
0.43
0.42
0.32
0.60
0.01
0.01
0.38
0.37
0.29
0.45
-0.39
0.57
0.52
0.45
-0.56
0.30
0.28
0.26
-0.23
-0.22
-0.22
0.42
0.36
0.31

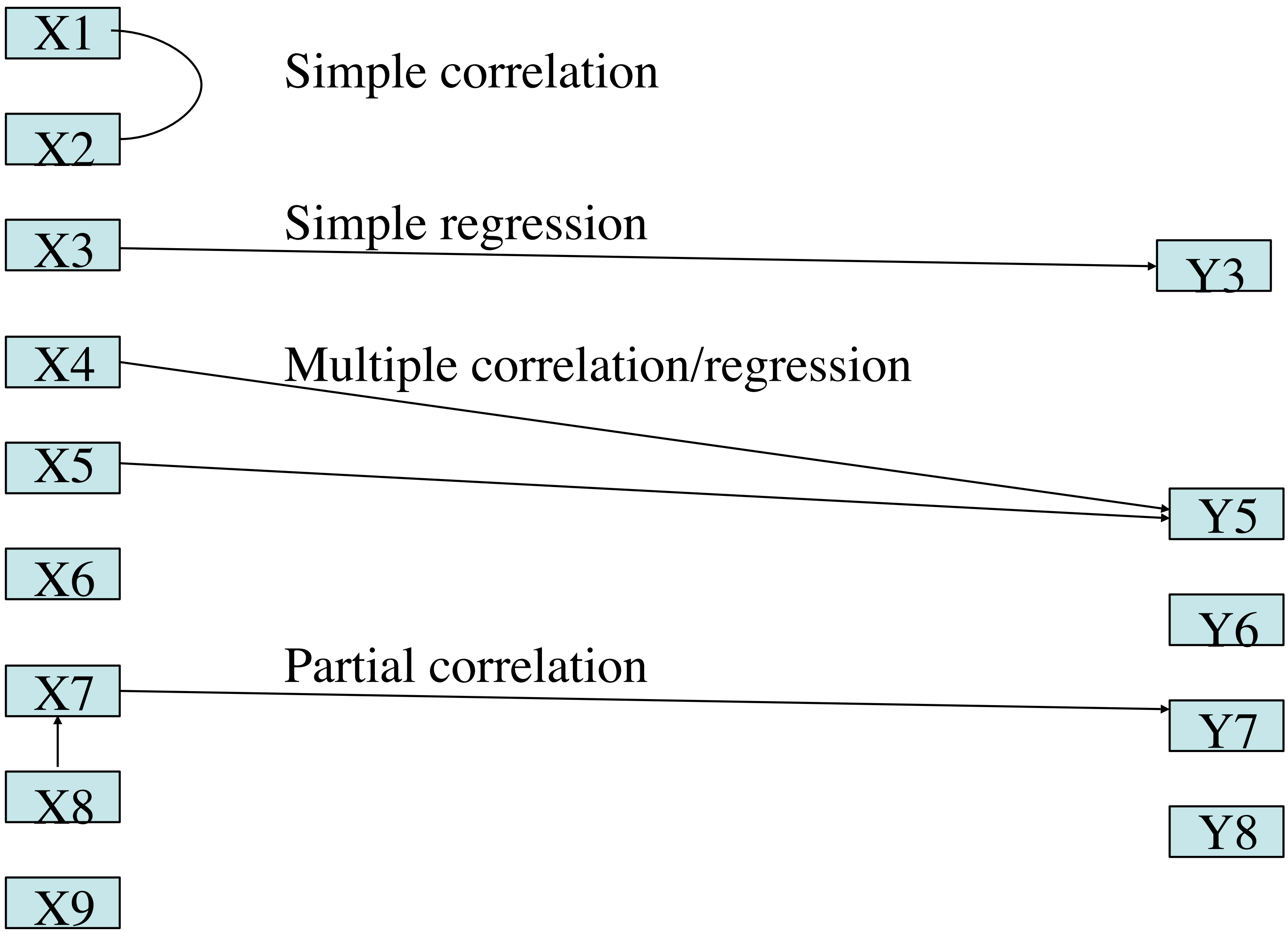
Psychometric Theory: A conceptual Syllabus



Measures of relationship

- Regression $y = bx + c$
  - $b_{y.x} = Cov_{xy} / Var_{x}$
- Correlation
  - $r_{xy} = Cov_{xy}/sqrt(V_x * V_y)$
  - Pearson Product moment correlation
    - Spearman (ppmc on ranks)
    - Point biserial (x is dichotomous, y continuous)
    - Phi (x, y both dichotomous)

Variance, Covariance, and Correlation



Measures of relationships with more than 2 variables

- Partial correlation
  - The relationship between x and y with z held constant (z removed)
- Multiple correlation
  - The relationship of x1 + x2 with y
  - Weight each variable by its independent contribution

Problems with correlations

- Simpson's paradox and the problem of aggregating groups
  - Within group relationships are not the same as between group or pooled relationships
- Phi coefficients and the problem of unequal marginals
- Alternative interpretations of partial correlations