Psychology 205: Research Methods in Psychology
Correlation and Regression

William Revelle

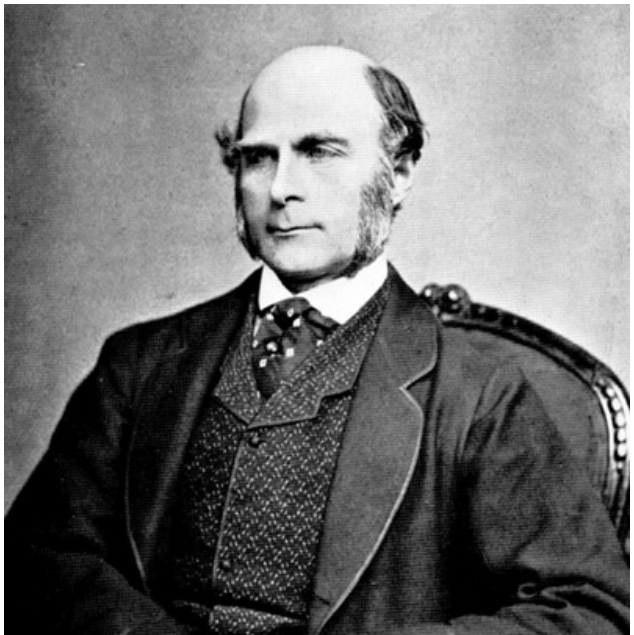
Department of Psychology
Northwestern University
Evanston, Illinois USA

October, 2014

Outline

Francis Galton 1822-1911

Francis Galton (1822-1911) was among the most influential psychologists of the 19th century. He did pioneering work on the correlation coefficient, behavior genetics and the measurement of individual differences. He introspectively examined the question of free will and introduced the lexical hypothesis to the study of personality and character. In addition to psychology, he did pioneering work in meteorology and introduced the scientific use of fingerprints. Whenever he could, he counted.



Karl Pearson 1857-1936

Carl (Karl) Pearson was among the most influential statisticians of the early 20th century. Founder of the statistics department at University College London. He developed the Pearson Product Moment Correlation Coefficient, its special case the ϕ coefficient, and the tetrachoric correlation. Major behavior geneticist and eugenicist.

Charles Spearman 1863-1945

Charles Spearman (1863-1945) was the leading psychometrician of the early 20th century. His work on the classical test theory, factor analysis, and the g theory of intelligence continues to influence psychometrics, statistics, and the study of intelligence. More than 100 years after their publication, his most influential papers remain two of the most frequently cited articles in psychometrics and intelligence.

Galton's height data

Table : The relationship between the average of both parents (mid parent) and the height of their children. The basic data table is from **?** who used these data to introduce "reversion to mediocrity" (the median). We now know this as regression to the mean (and thus, linear regression). The data are available as part of the **UsingR** or **psych** packages.

```
> library(psych)
> data(galton)
> galton.tab <- table(galton)
> galton.tab[order(rank(rownames(galton.tab)),decreasing=TRUE),] #sort it by decreasing row va
```

	child												
parent	61.7	62.2	63.2	64.2	65.2	66.2	67.2	68.2	69.2	70.2	71.2	72.2	73.2	73.7
73	0	0	0	0	0	0	0	0	0	0	0	1	3	0
72.5	0	0	0	0	0	0	0	1	2	1	2	7	2	4
71.5	0	0	0	0	1	3	4	3	5	10	4	9	2	2
70.5	1	0	1	0	1	1	3	12	18	14	7	4	3	3
69.5	0	0	1	16	4	17	27	20	33	25	20	11	4	5
68.5	1	0	7	11	16	25	31	34	48	21	18	4	3	0
67.5	0	3	5	14	15	36	38	28	38	19	11	4	0	0
66.5	0	3	3	5	2	17	17	14	13	4	0	0	0	0
65.5	1	0	9	5	7	11	11	7	7	5	2	1	0	0
64.5	1	1	4	4	1	5	5	0	2	0	0	0	0	0
64	1	0	2	4	1	2	2	1	1	0	0	0	0	0

Galton's height data

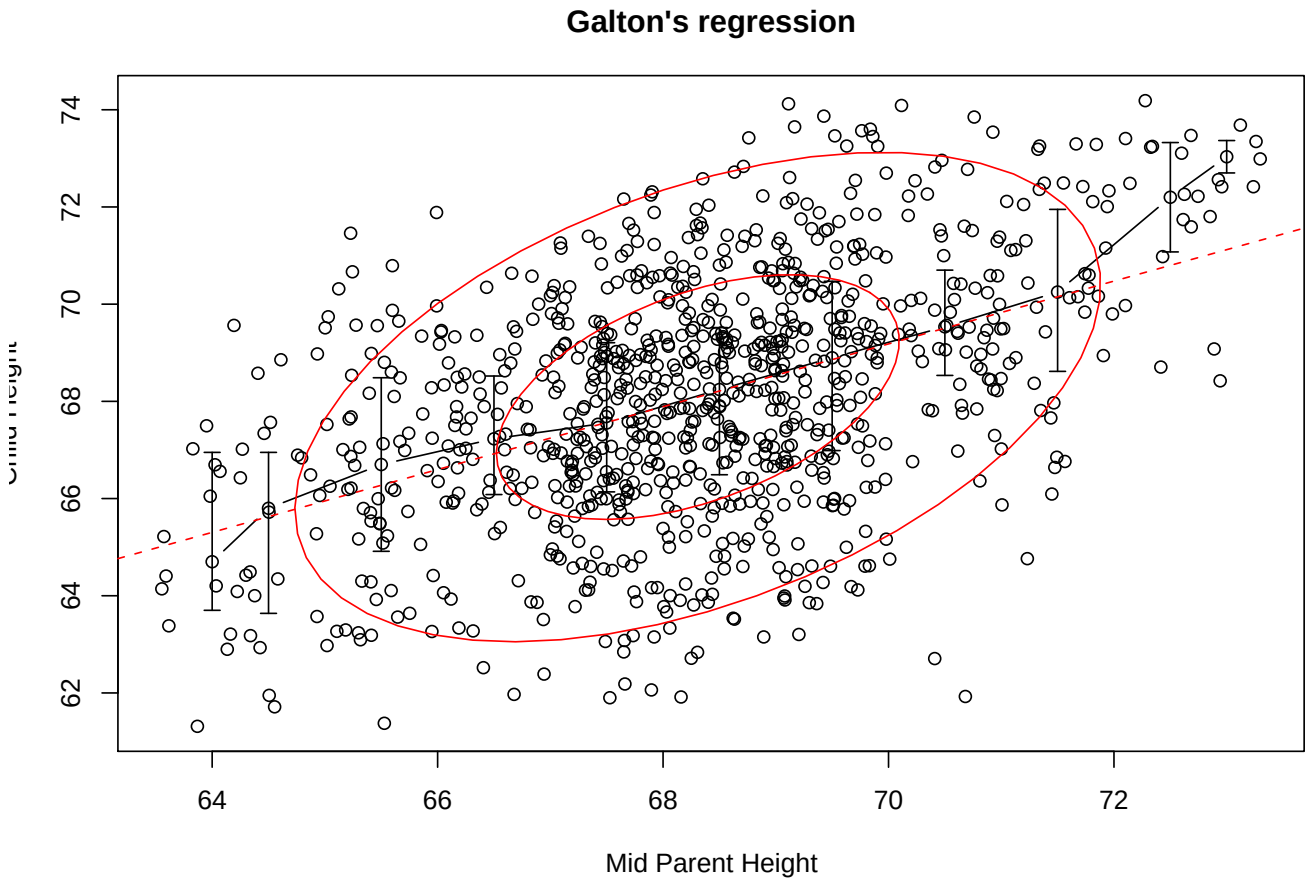


Figure : Galton's data can be plotted to show the relationships between mid parent and child heights. Because the original data are grouped, the data points have been *jittered* to emphasize the density of points along the median. The bars connect the first, 2nd (median) and third quartiles. The dashed line is the best fitting linear fit, the ellipses represent one and two standard deviations from the mean.

Plotting Galton's Height data

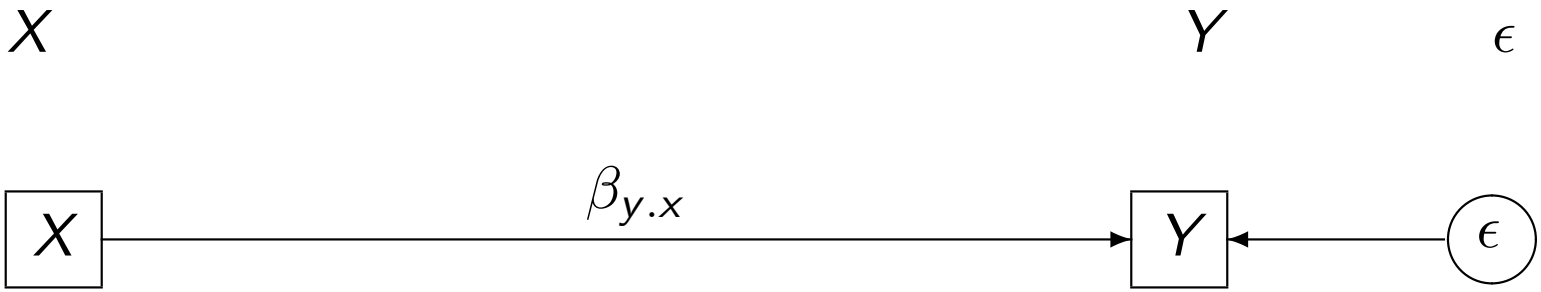
This makes use of a number of different functions, some from Base R, some from the *psych*.

R code

```
library(psych)  #Make psych active
data(galton)     #get the data
#this next line does most of the work
#note that we jitter the data points
plot(jitter(galton$parent,factor=5),jitter(galton$child,factor=5),
     xlab="Mid Parent Height",ylab="Child Height",
     main="Galton's regression")
interp.qplot.by(galton$child,galton$parent,add=TRUE)
ellipses(galton$parent,galton$child,data=FALSE,
     smooth=FALSE,add=TRUE,lm=TRUE)
```

See the help file on interp.qplot.by and ellipses

Bivariate Regression – Predicting Y as a function of X

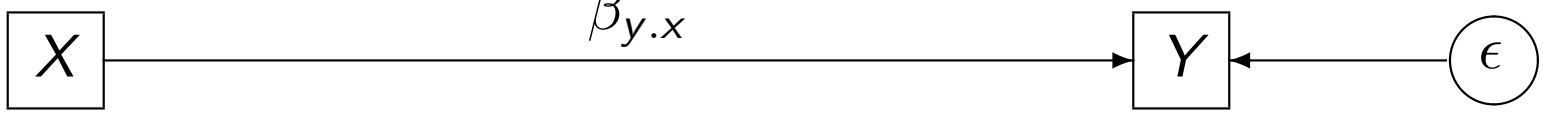


$$\hat{y} = \beta_{y.x} x + \epsilon$$

$$\beta_{y.x} = \frac{\sigma_{xy}}{\sigma_x^2}$$

Bivariate Regression – Predicting X as a function of Y

δ X Y ϵ



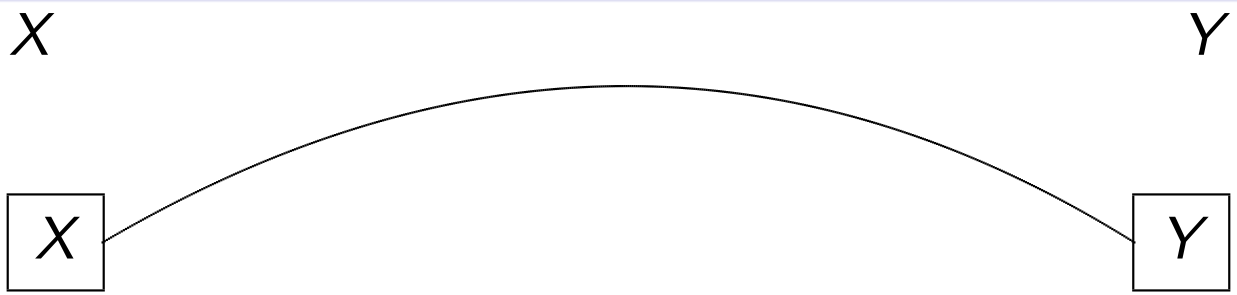
$$\hat{y} = \beta_{y.x}x + \epsilon$$

$$\beta_{y.x} = \frac{\sigma_{xy}}{\sigma_x^2}$$

$$\hat{x} = \beta_{x.y}y + \delta$$

$$\beta_{y.x} = \frac{\sigma_{xy}}{\sigma_y^2}$$

Bivariate Correlation is the geometric average of the two regressions



$$\hat{x} = \beta_{x.y} y + \delta \qquad \hat{y} = \beta_{y.x} x + \epsilon$$

$$\beta_{y.x} = \frac{\sigma_{xy}}{\sigma_y^2} \qquad \beta_{y.x} = \frac{\sigma_{xy}}{\sigma_x^2}$$

$$r_{xy} = \frac{\sigma_{xy}}{\sqrt{\sigma_x^2 \sigma_y^2}}$$

The variance and the variance of a composite (advanced)

1. If $\mathbf{x_1}$ and $\mathbf{x_2}$ are column vectors of N observations centered around their mean (that is, deviation scores) their variances are $V_{x1} = \sum x_{i1}^2/(N-1)$ and $V_{x2} = \sum x_{i2}^2/(N-1)$, or, in matrix terms $V_{\mathbf{x}1} = \mathbf{x_1}\mathbf{x_1}'/(N-1)$ and $V_{\mathbf{x}2} = \mathbf{x_2}\mathbf{x_2}'/(N-1)$.
2. The variance of the composite made up of the sum of the corresponding scores, $\mathbf{x} + \mathbf{y}$ is just

$$V_{(\mathbf{x1}+\mathbf{x2})} = \frac{\sum (x_i + y_i)^2}{N-1} = \frac{\sum x_i^2 + \sum y_i^2 + 2\sum x_i y_i}{N-1} = \frac{(\mathbf{x}+\mathbf{y})(\mathbf{x}+\mathbf{y})'}{N-1}. \quad (1)$$

Or, more generally,

$$\mathbf{S} = \begin{pmatrix} v_{x1} & c_{x1x2} & \cdots & c_{x1xn} \\ c_{x1x2} & v_{x2} & & c_{x2xn} \\ \vdots & & \ddots & \vdots \\ c_{x1xn} & c_{x2xn} & \cdots & v_{xn} \end{pmatrix}$$

Sums as matrix products (advanced)

$$V_{\mathbf{X}} = \sum \frac{\mathbf{X}\mathbf{X}'}{N-1} = \frac{\mathbf{1}'(\mathbf{X}\mathbf{X}')\mathbf{1}}{N-1}.$$

$$V_{\mathbf{Y}} = \sum \frac{\mathbf{Y}\mathbf{Y}'}{N-1} = \frac{\mathbf{1}'(\mathbf{Y}\mathbf{Y}')\mathbf{1}}{N-1}$$

and

$$C_{\mathbf{XY}} = \sum \frac{\mathbf{X}\mathbf{Y}'}{N-1} = \frac{\mathbf{1}'(\mathbf{X}\mathbf{Y}')\mathbf{1}}{N-1}$$

Use R

Get the data from a remote data source

A nice feature of R is that you can read from remote data sets. The example dataset is on the personality-project.org server. Get it and describe it.

R code

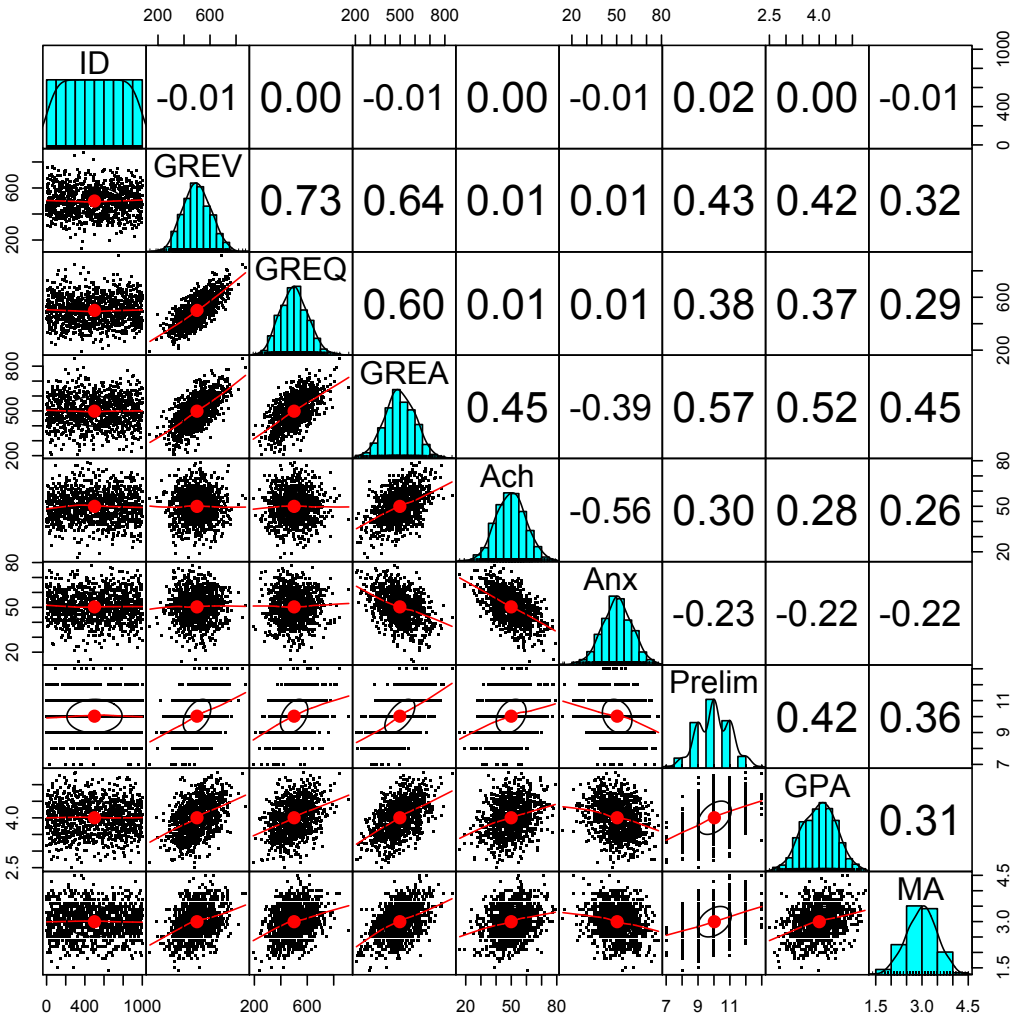
```
>  datafilename=
"http://personality-project.org/r/datasets/psychometrics.prob2.txt"
> mydata =read.table(datafilename,header=TRUE) #read the data file
> describe(mydata,skew=FALSE)
```

```
        var    n   mean     sd median trimmed    mad   min     max  range   se
ID        1 1000 500.50 288.82 500.50  500.50 370.65   1.0 1000.00 999.00 9.13
GREV      2 1000 499.77 106.11 497.50  498.75 106.01 138.0  873.00 735.00 3.36
GREQ      3 1000 500.53 103.85 498.00  498.51 105.26 191.0  914.00 723.00 3.28
GREA      4 1000 498.13 100.45 495.00  498.67  99.33 207.0  848.00 641.00 3.18
Ach       5 1000  49.93   9.84  50.00   49.88  10.38  16.0   79.00  63.00 0.31
Anx       6 1000  50.32   9.91  50.00   50.43  10.38  14.0   78.00  64.00 0.31
Prelim    7 1000  10.03   1.06  10.00   10.02   1.48   7.0   13.00   6.00 0.03
GPA       8 1000   4.00   0.50   4.02    4.01   0.53   2.5    5.38   2.88 0.02
MA        9 1000   3.00   0.49   3.00    3.00   0.44   1.4    4.50   3.10 0.02
```

Plot it using the `pairs.panels` function.

Use the pairs.panels function to show a splom plot (use gap=0 and pch='.').

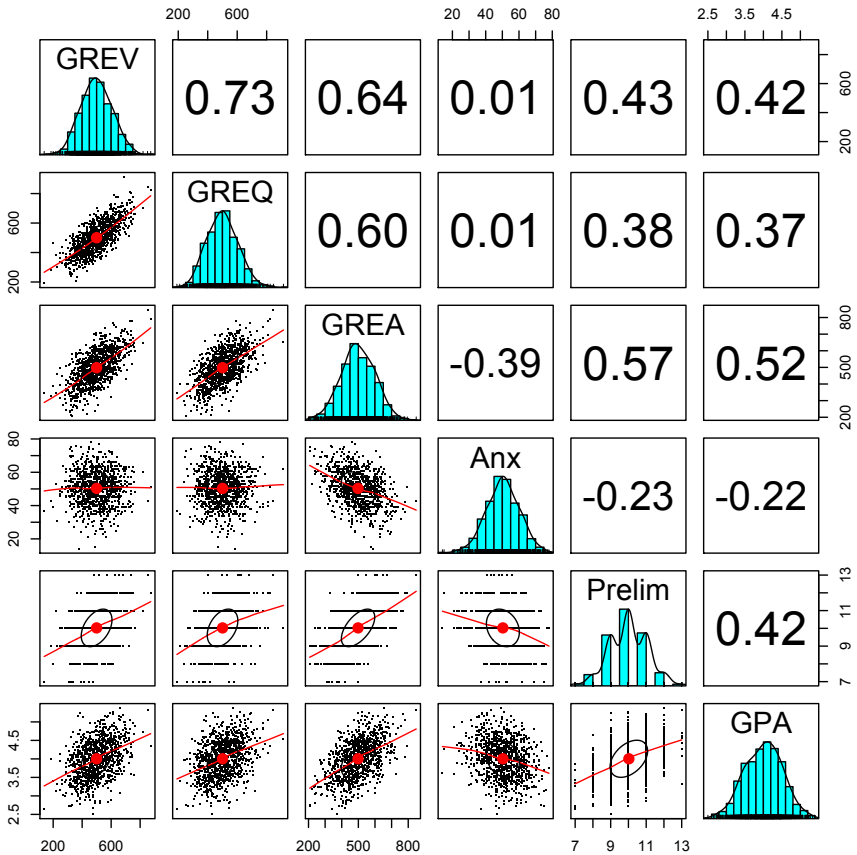
>pairs.panels(mydata,pch=".",gap=0) #pch='.' makes for a cleaner plot



Plot a subset of the data using the c() function (concatenate).

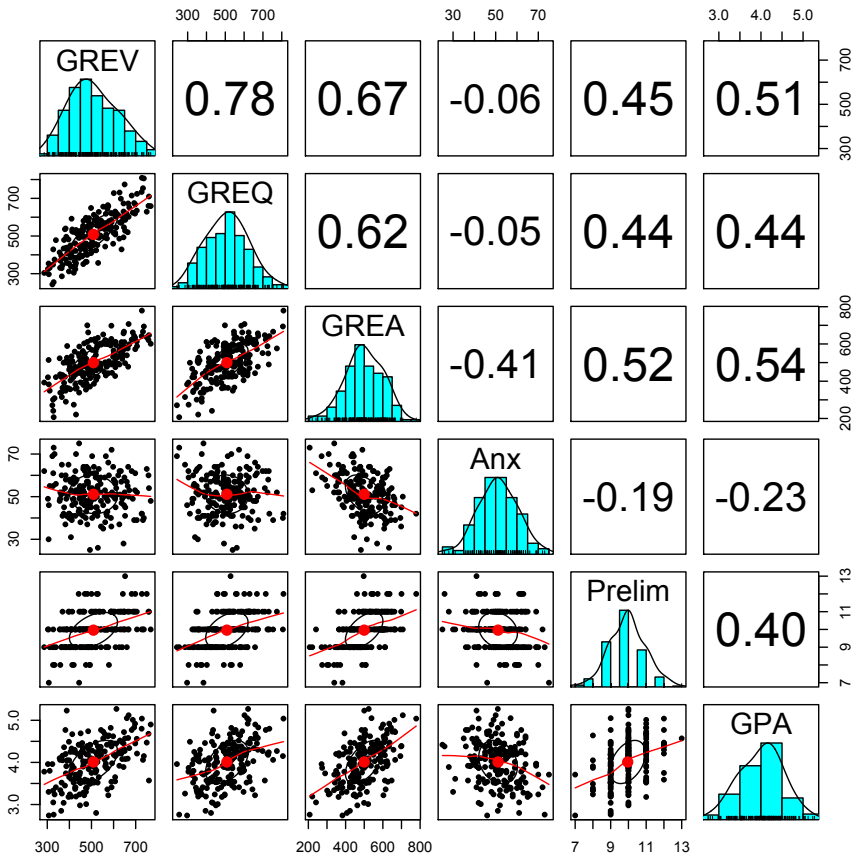
Use the pairs.panels function to show a splom plot. Select a subset of variables using the c() function.

>pairs.panels(mydata[c(2:4,6:8)],pch='.')



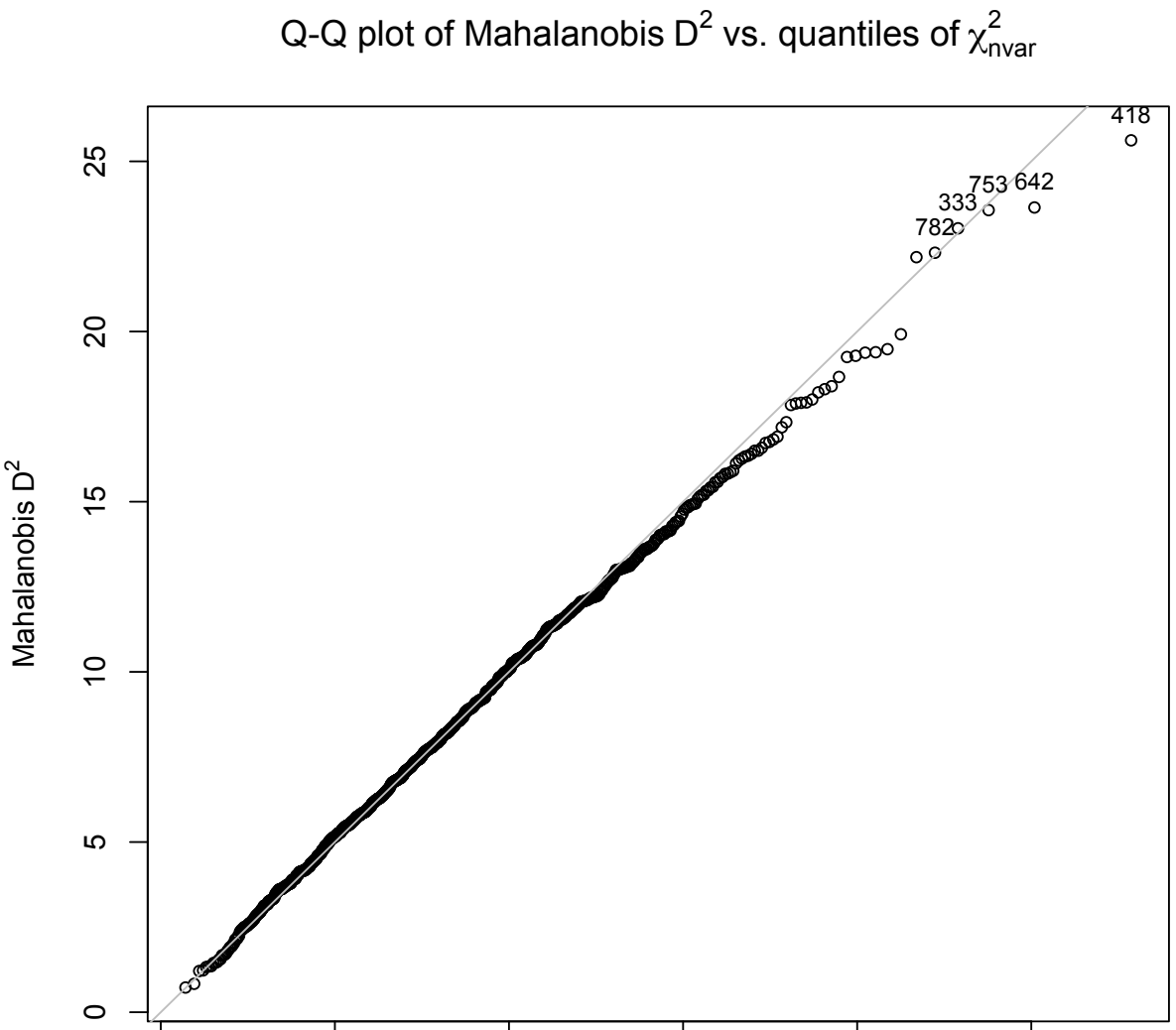
Do this for the first 200 subjects

```
> pairs.panels(mydata[mydata$ID < 200,c(2:4,6:8)])
```



Look for outliers by comparing the Mahalanobis d^2 to the expected values

```
d2 <- outlier(mydata[-1],cex=.8)
```



0 center the data

In order to do interaction terms in regressions, it is necessary to 0 center the data. We need to turn the result into a data.frame in order to use it in the regression function.

```
> cent <- data.frame(scale(mydata,scale=FALSE))
> describe(cent,skew=FALSE)
```

```
         var    n mean     sd median trimmed    mad     min    max  range   se
ID         1 1000    0 288.82   0.00    0.00 370.65 -499.50 499.50 999.00 9.13
GREV       2 1000    0 106.11  -2.27   -1.02 106.01 -361.77 373.23 735.00 3.36
GREQ       3 1000    0 103.85  -2.53   -2.02 105.26 -309.53 413.47 723.00 3.28
GREA       4 1000    0 100.45  -3.13    0.54  99.33 -291.13 349.87 641.00 3.18
Ach        5 1000    0   9.84   0.07   -0.05  10.38  -33.93  29.07  63.00 0.31
Anx        6 1000    0   9.91  -0.32    0.11  10.38  -36.32  27.68  64.00 0.31
Prelim     7 1000    0   1.06  -0.03    0.00   1.48   -3.03   2.97   6.00 0.03
GPA        8 1000    0   0.50   0.02    0.00   0.53   -1.50   1.38   2.88 0.02
MA         9 1000    0   0.49   0.00    0.00   0.44   -1.60   1.50   3.10 0.02
```

The standard deviations and ranges have not changed. However, the means are all 0. We use the `scale` function with the scale=FALSE option.

The standardized data

Alternatively, we could standardize it to make the variances all equal as well.

```
> z.data <- data.frame(scale(my.data))
>  describe(z.data)
        var    n mean sd median trimmed  mad   min  max range  skew kurtosis   se
ID        1 1000    0  1   0.00    0.00 1.28 -1.73 1.73  3.46  0.00    -1.20 0.03
GREV      2 1000    0  1  -0.02   -0.01 1.00 -3.41 3.52  6.93  0.09    -0.07 0.03
GREQ      3 1000    0  1  -0.02   -0.02 1.01 -2.98 3.98  6.96  0.22     0.08 0.03
GREA      4 1000    0  1  -0.03    0.01 0.99 -2.90 3.48  6.38 -0.02    -0.06 0.03
Ach       5 1000    0  1   0.01   -0.01 1.05 -3.45 2.95  6.40  0.00     0.02 0.03
Anx       6 1000    0  1  -0.03    0.01 1.05 -3.67 2.79  6.46 -0.14     0.14 0.03
Prelim    7 1000    0  1  -0.02    0.00 1.40 -2.86 2.81  5.67 -0.02    -0.01 0.03
GPA       8 1000    0  1   0.03    0.01 1.06 -3.00 2.74  5.74 -0.07    -0.29 0.03
MA        9 1000    0  1   0.01    0.01 0.90 -3.23 3.04  6.27 -0.07    -0.09 0.03
```

We can standardize the data by subtracting the mean and dividing though by the standard deviation. We use the `scale` function to do this for us.

Show how the correlations do not change with standardization

Find the correlations using the `lowerCor` function. This, by default, uses *pairwise* Pearson correlations and rounds to two decimals. Compare with the standard `cor` function. This will not work with missing data unless specified as use="pairwise"

```
> lowerCor(my.data)
```

```
        ID    GREV  GREQ  GREA  Ach   Anx   Prelm GPA   MA
ID       1.00
GREV    -0.01  1.00
GREQ     0.00  0.73  1.00
GREA    -0.01  0.64  0.60  1.00
Ach      0.00  0.01  0.01  0.45  1.00
Anx     -0.01  0.01  0.01 -0.39 -0.56  1.00
Prelim   0.02  0.43  0.38  0.57  0.30 -0.23  1.00
GPA      0.00  0.42  0.37  0.52  0.28 -0.22  0.42  1.00
MA      -0.01  0.32  0.29  0.45  0.26 -0.22  0.36  0.31  1.00
```

```
> lowerCor(z.data)
```

```
        ID    GREV  GREQ  GREA  Ach   Anx   Prelm
ID       1.00
GREV    -0.01  1.00
GREQ     0.00  0.73  1.00
GREA    -0.01  0.64  0.60  1.00
Ach      0.00  0.01  0.01  0.45  1.00
Anx     -0.01  0.01  0.01 -0.39 -0.56  1.00
Prelim   0.02  0.43  0.38  0.57  0.30 -0.23  1.00
GPA      0.00  0.42  0.37  0.52  0.28 -0.22  0.42
MA      -0.01  0.32  0.29  0.45  0.26 -0.22  0.36
```

Show that the two matrices do not differ using the lowerUpper function

```
r <- lowerCor(my.data)  #find the original correlations
z <- lowerCor(z.data)   #find the z transformed correlations
lu <- lowerUpper(r,z,diff=TRUE) #combine into one matrix
           and take the difference

round(lu,2)  #round to 2 decimals
#this is the same as
round(lu,digits=2)  #parameters can be spelled out,
    or defined by position
```

```
          ID GREV GREQ  GREA   Ach   Anx Prelim  GPA MA
ID        NA 0.00 0.00  0.00  0.00  0.00   0.00 0.00  0
GREV   -0.01   NA 0.00  0.00  0.00  0.00   0.00 0.00  0
GREQ    0.00 0.73   NA  0.00  0.00  0.00   0.00 0.00  0
GREA   -0.01 0.64 0.60    NA  0.00  0.00   0.00 0.00  0
Ach     0.00 0.01 0.01  0.45    NA  0.00   0.00 0.00  0
Anx    -0.01 0.01 0.01 -0.39 -0.56    NA   0.00 0.00  0
Prelim  0.02 0.43 0.38  0.57  0.30 -0.23     NA 0.00  0
GPA     0.00 0.42 0.37  0.52  0.28 -0.22   0.42   NA  0
MA     -0.01 0.32 0.29  0.45  0.26 -0.22   0.36 0.31 NA
```

Lets get some real data (from a remote server)

1. Data are from graduate applicants to NU psychology
   - Graduate Record Verbal, Quant, Advanced
   - GPA
   - The "Oregon" ranking system $\frac{GREV+GREQ}{200} + GPA$
2. Data are on the personality-project server at

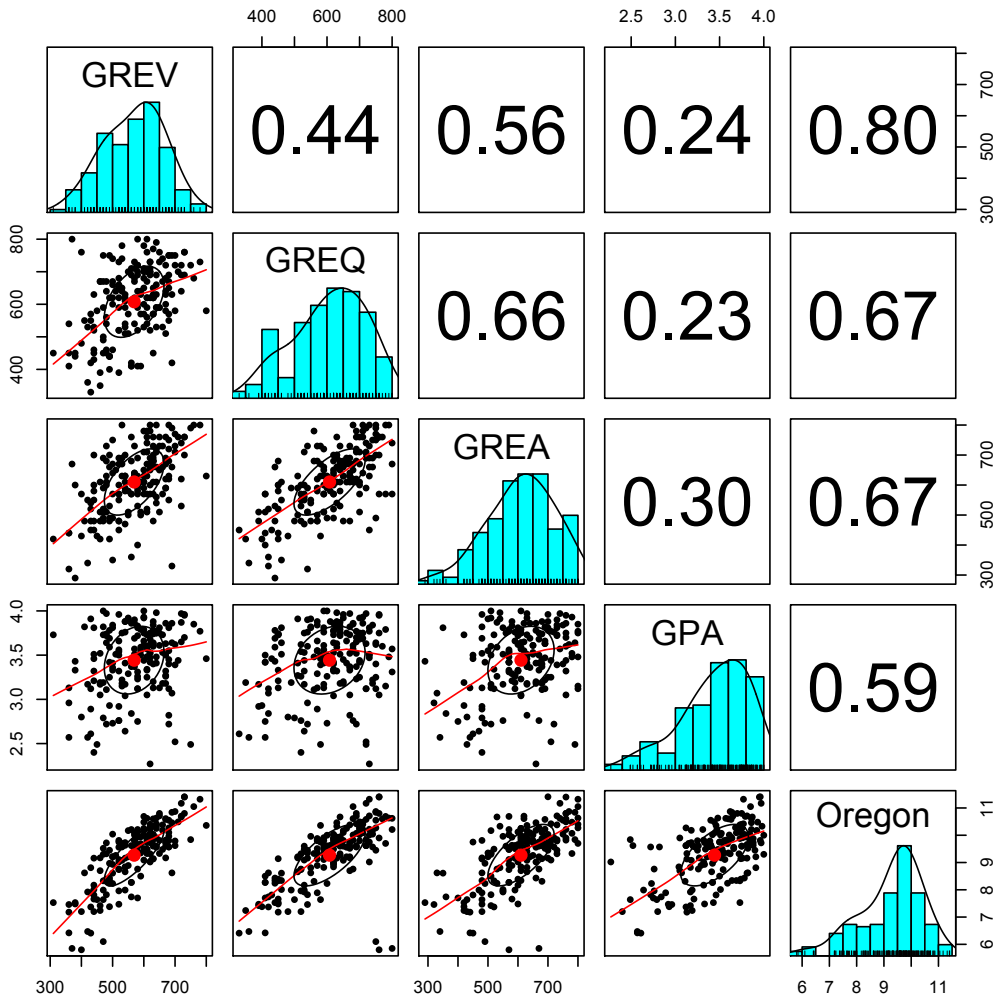
   "http://personality-project.org/r/tutorials/summerschool.14/applicants.csv"
3. Give the file name of the data to get and then read the data, specifying that we have headers for the columns.
4. Find basic descriptive statistics, then show the graphics, look for outliers

R code

```
> fn <-
"http://personality-project.org/r/tutorials/summerschool.14/applicants.csv"
> grad <- read.table(fn,header=TRUE,sep=",")
> bad <- outlier(grad,cex=.8)
> pairs.panels(grad,bg=c("yellow","blue")[(bad > 40)+1],pch=21)
> pairs.panels(grad[(bad < 40),])  #dump the outliers
> pairs.panels(gradf[gradf[5]>10,])  #select just the high Oregon scores
```

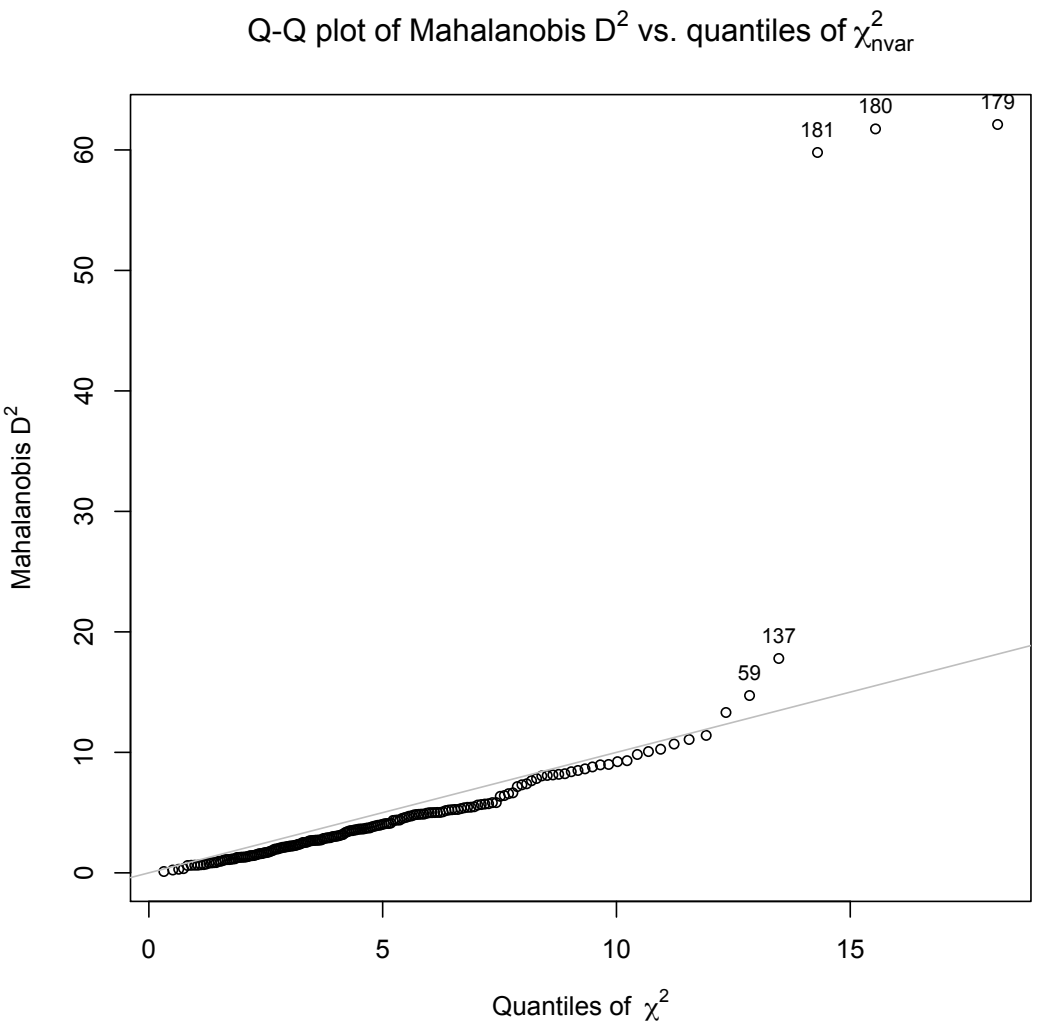
Scatter Plot Matrix showing correlation and LOESS regression

Real data taken from NU applicants.



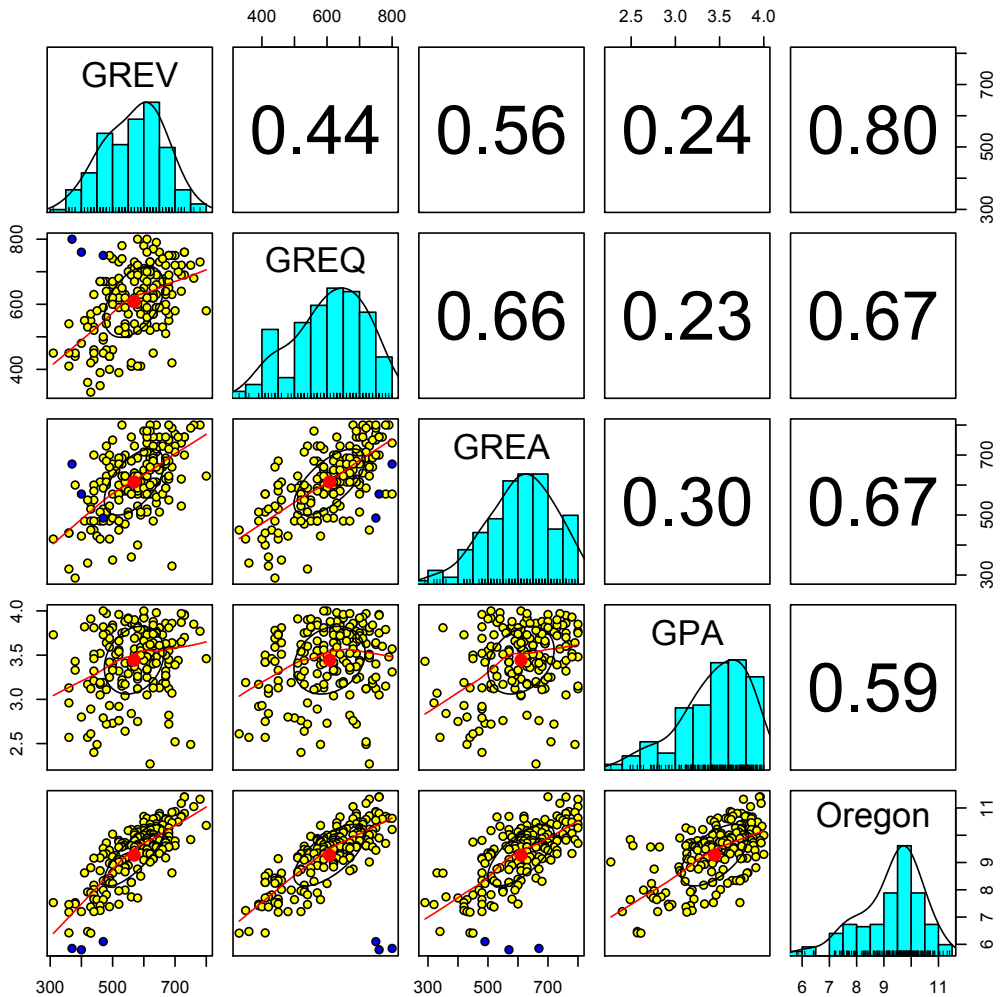
outlier detection examines the comparing the Mahalanobis d^2 to the expected values

Real data taken from NU applicants.



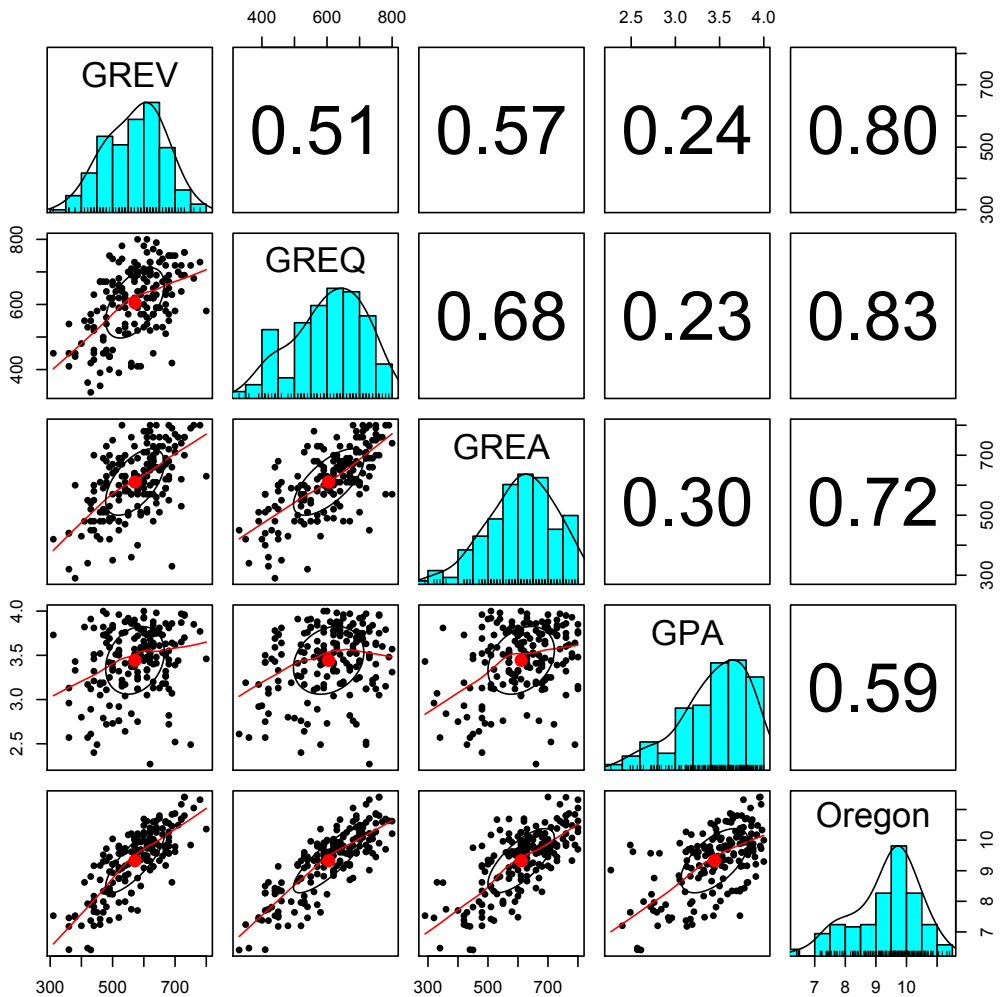
SPLOM showing the most extreme d^2 cases

Real data taken from NU applicants.

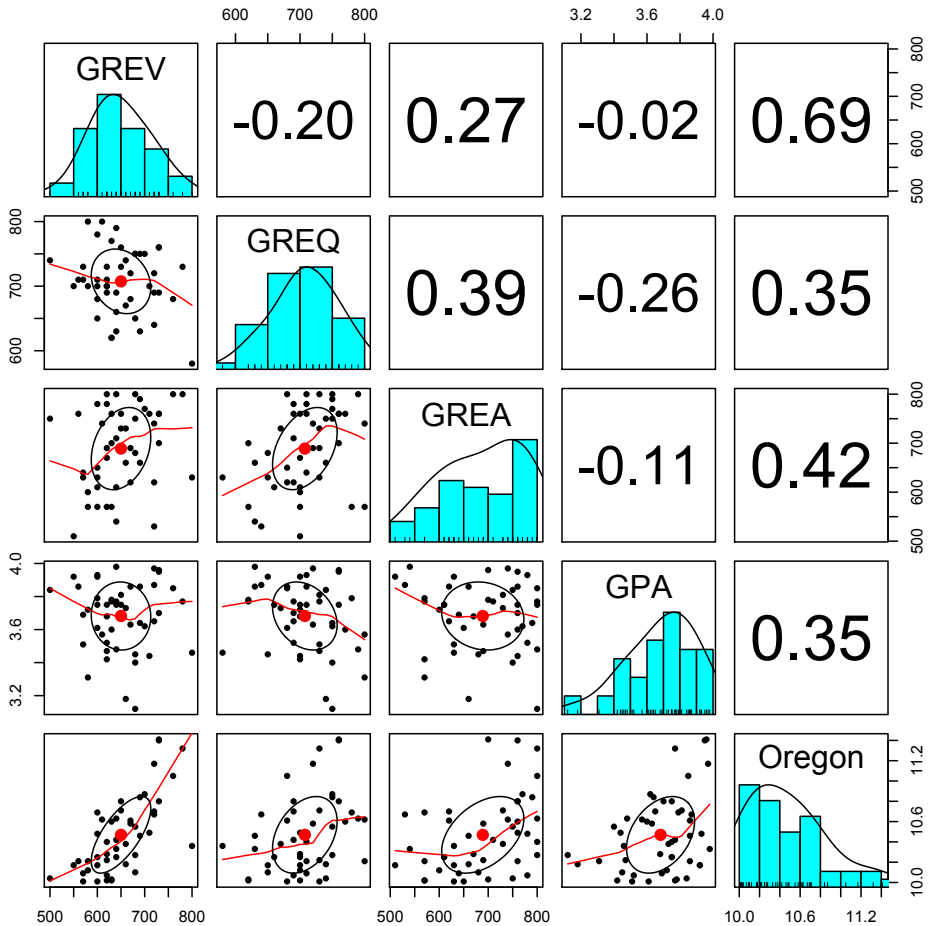


Scatter Plot Matrix showing correlation and LOESS regression – outliers removed

Real data taken from NU applicants.

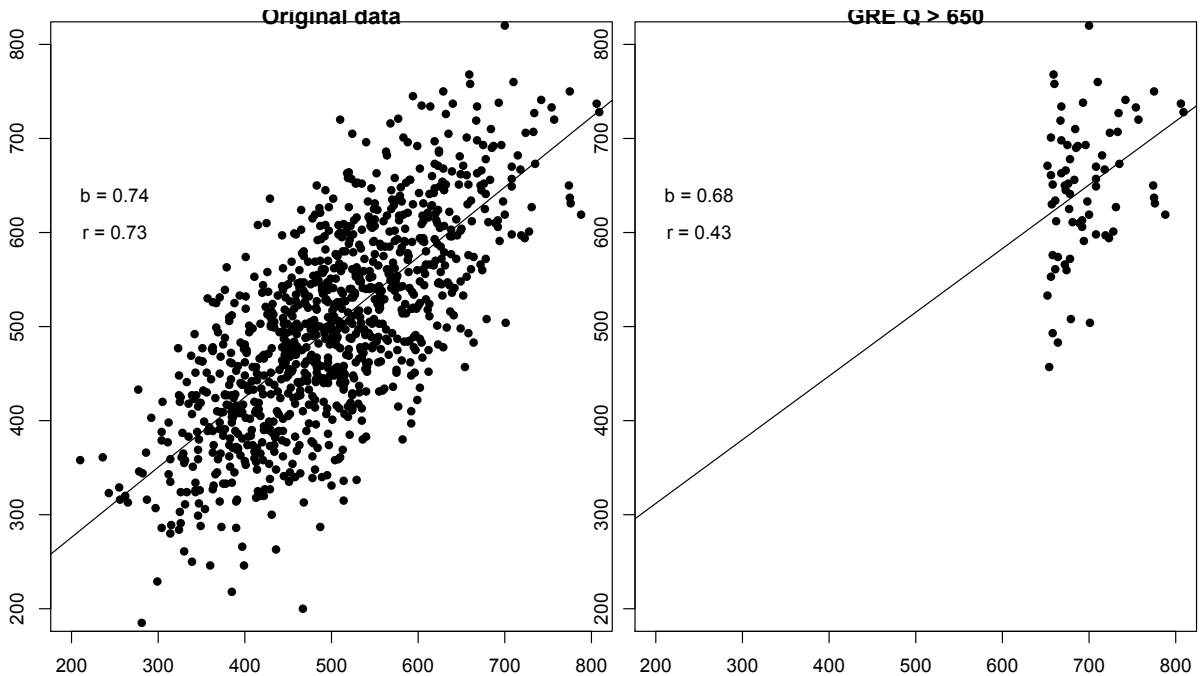


The effect of selection on the correlation



- Consider what happens if we select a subset of the applicants according to linear sum of the three predictors:
  - The "Oregon" model
  - (GPA + (V+Q)/200) $>$ 10.0
- The range is truncated, but even more important, by using a compensatory selection model, we have changed the sign of the correlations.

Regression and restriction of range - Bivariate Normal data



Although the correlation is very sensitive, regression slopes are relatively insensitive to restriction of range.

R code for regression figures. We use the `with` and `paste` constructs.

R code

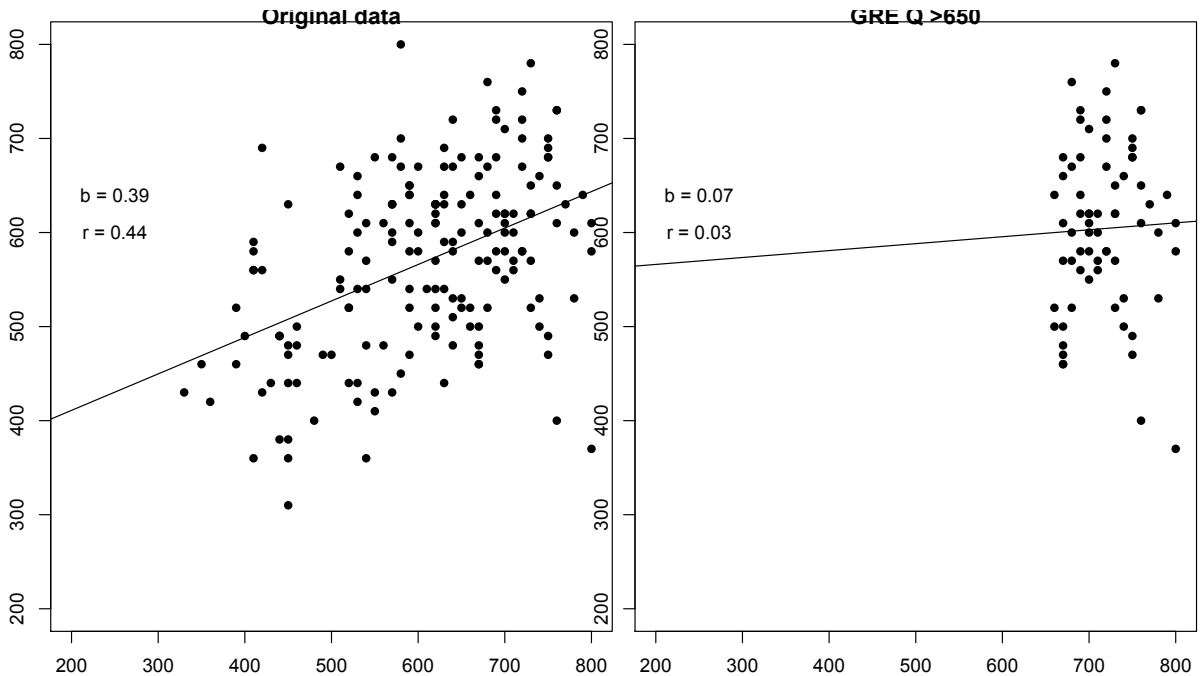
```
op <- par(mfrow=c(1,2))  #two panel graph
 with(mydata,{
 b = round(lm(GREV ~ GREQ)$coeff[2],2) #find the slope
 r = round(cor(GREV, GREQ),2)  #find the correlation
 plot(GREV ~ GREQ,ylim=c(200,800),xlim=c(200,800),
            main='Original data', pch=16)
 abline(lm(GREV ~ GREQ)  )  #draw the line
 text(250,600,paste('r =', r))   #add a label
 text(250,640,paste('b =', b)) #add another label
 })  #label it

 with(mydata[mydata$GREQ > 650,],{
  b = round(lm(GREV ~ GREQ)$coeff[2],2)
 r = round(cor(GREV, GREQ),2)
 plot(GREV ~ GREQ,ylim=c(200,800),xlim=c(200,800),
        main='GRE Q > 650',pch=16)
 abline(lm(GREV ~ GREQ))
  text(250,600,paste('r =', r))
 text(250,640,paste('b =', b))
 })  #label it
 op <- par(mfrow=c(1,1)) #switch back to one panel
```

Problems with restriction of range

1. In the normal case, where the restriction is externally applied
   - The slope does not change very much
   - The correlation will change
2. But, in the case of students applying to graduate school, they are implicitly building in a compensatory model.
   - The very high Verbal and high quantitative people are applying to different fields?
   - This leads to an interesting problem, in that selecting on one extreme is implicitly selecting with a compensatory model.

Regression and restriction of range - Self selection compensation case



The correlation is very sensitive, as is the regression slope when self compensation is occurring.

R code for regression figures. We use the `with` and `paste` constructs.

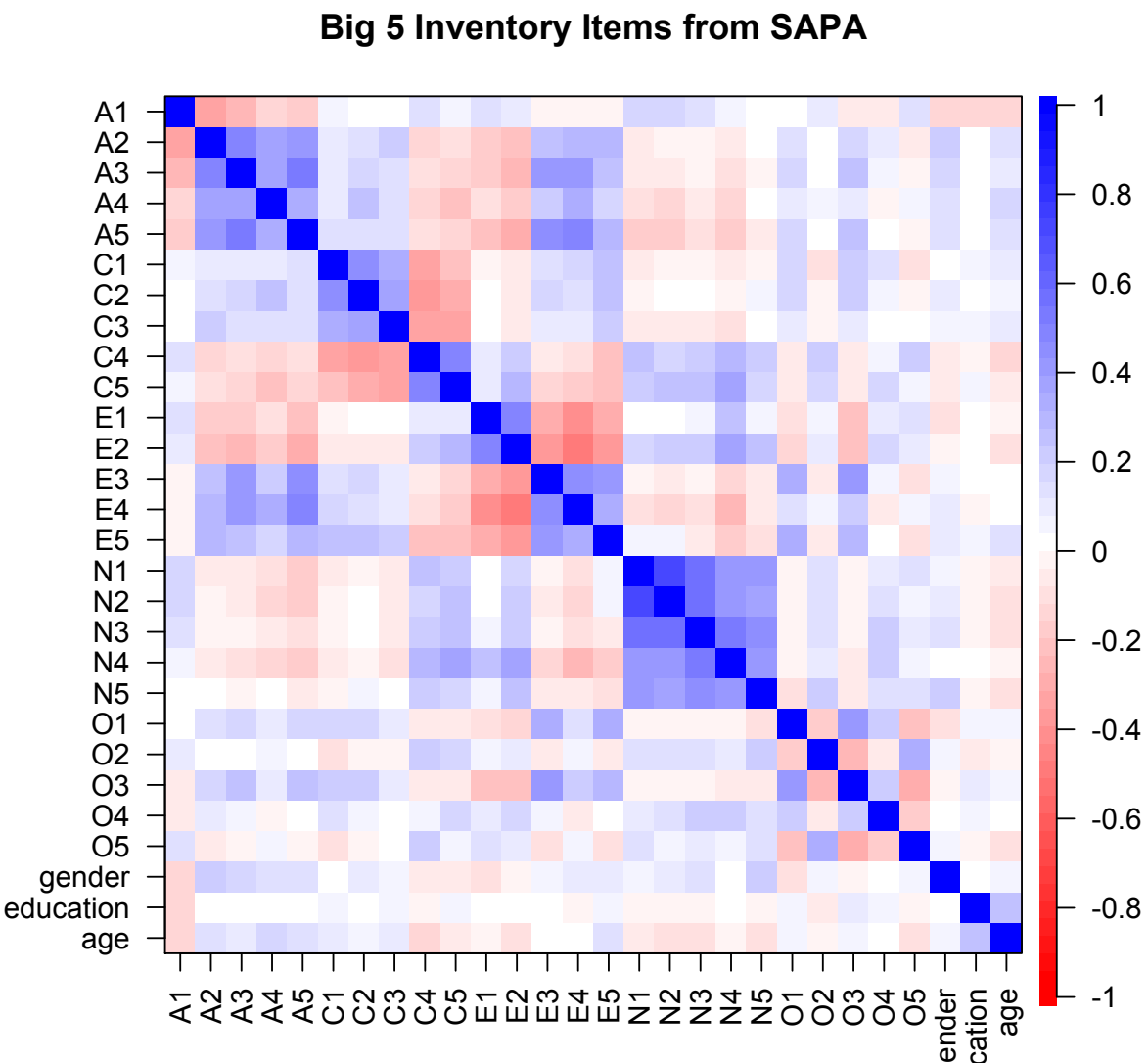
R code

```
op <- par(mfrow=c(1,2))  #two panel graph
 with(grad,{
 b = round(lm(GREV ~ GREQ)$coeff[2],2) #find the slope
 r = round(cor(GREV, GREQ,use='pairwise'),2)  #find the correlation
 plot(GREV ~ GREQ,ylim=c(200,800),xlim=c(200,800),
            main='Original data', pch=16)
 abline(lm(GREV ~ GREQ)  )  #draw the line
 text(250,600,paste('r =', r))   #add a label
 text(250,640,paste('b =', b)) #add another label
 })  #label it

 with(grad[grad$GREQ > 650,],{
  b = round(lm(GREV ~ GREQ)$coeff[2],2)
 r = round(cor(GREV, GREQ),2)
 plot(GREV ~ GREQ,ylim=c(200,800),xlim=c(200,800),
        main='GRE Q >650',pch=16)
 abline(lm(GREV ~ GREQ))
  text(250,600,paste('r =', r))
 text(250,640,paste('b =', b))
 })  #label it
 op <- par(mfrow=c(1,1)) #switch back to one panel
```

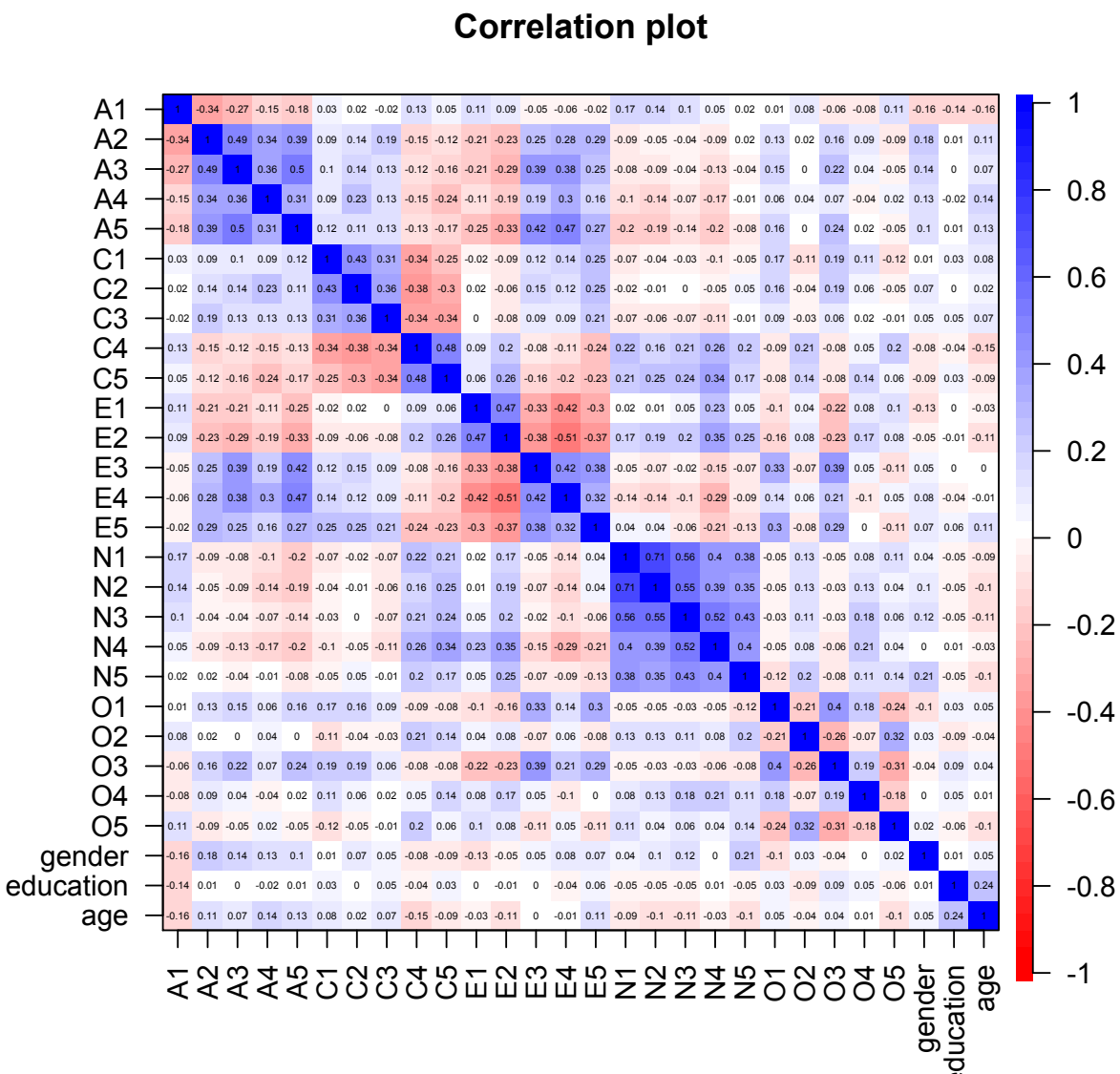
Show many correlations with a heat map using `cor.plot`.

```
r <- lowerCor(bfi)  #find the correlations
 cor.plot(r) #show the correlations as a heat map
```



Show many correlations with a heat map using `cor.plot`.

```
cor.plot(r,numbers=TRUE) #show the correlations as a heat map
```



Alternative versions of the correlation coefficient

Table : A number of correlations are Pearson r in different forms, or with particular assumptions. If $r = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$, then depending upon the type of data being analyzed, a variety of correlations are found.

Coefficient	symbol	X	Y	Assumptions
Pearson	r	continuous	continuous	
Spearman	rho (ρ)	ranks	ranks	
Point bi-serial	r_{pb}	dichotomous	continuous	
Phi	ϕ	dichotomous	dichotomous	
Bi-serial	r_{bis}	dichotomous	continuous	normality
Tetrachoric	r_{tet}	dichotomous	dichotomous	bivariate normality
Polychoric	r_{pc}	categorical	categorical	bivariate normality

The ϕ coefficient is just a Pearson r on dichotomous data

Table : The basic table for a phi, ϕ coefficient, expressed in raw frequencies in a four fold table is taken from **?**

	Success	Failure	Total
Accept	A	B	$R_1 = A + B$
Reject	C	D	$R_2 = C + D$
Total	$C_1 = A + C$	C_2=B + D	n = A + B + C + D

In terms of the raw data coded 0 or 1, the *phi coefficient* can be derived directly by direct substitution, recognizing that the only non zero product is found in the A cell

$$n\sum X_i Y_i - \sum X_i \sum Y_i = nA - R_1 C_1$$

$$\phi = \frac{AD - BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}} = \frac{A - R_1 C_1}{\sqrt{R_1 R_2 C_1 C_2}} \tag{2}$$

Correlation size $\neq$ causal importance

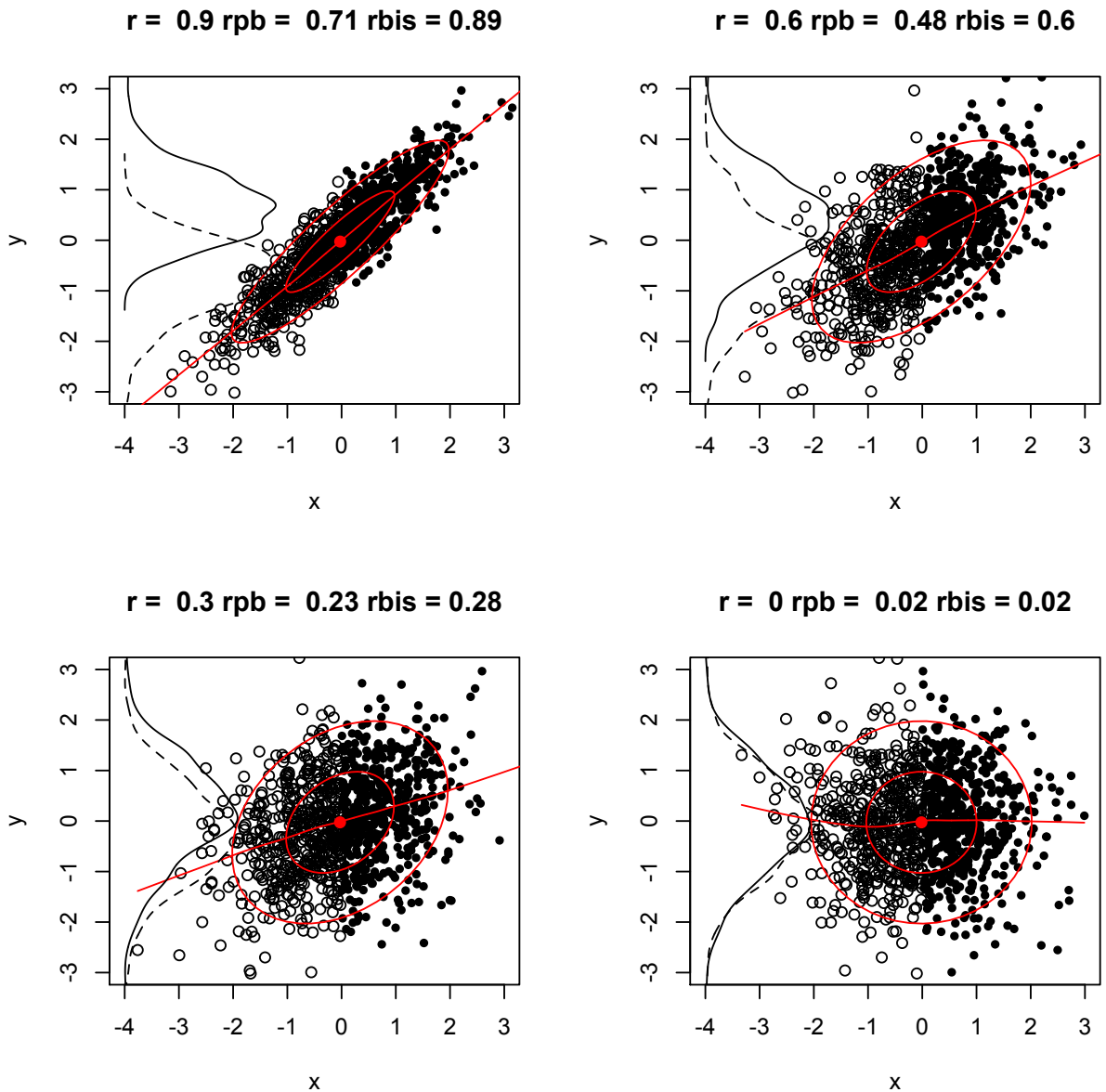
Table : The relationship between sex and pregnancy (hypothetical data)

	Pregnant	Not Pregnant	Total
Intercourse	2	1,041	1,043
No intercourse	0	6,257	6,257
Total	2	7,298	7,300
Phi	.04		

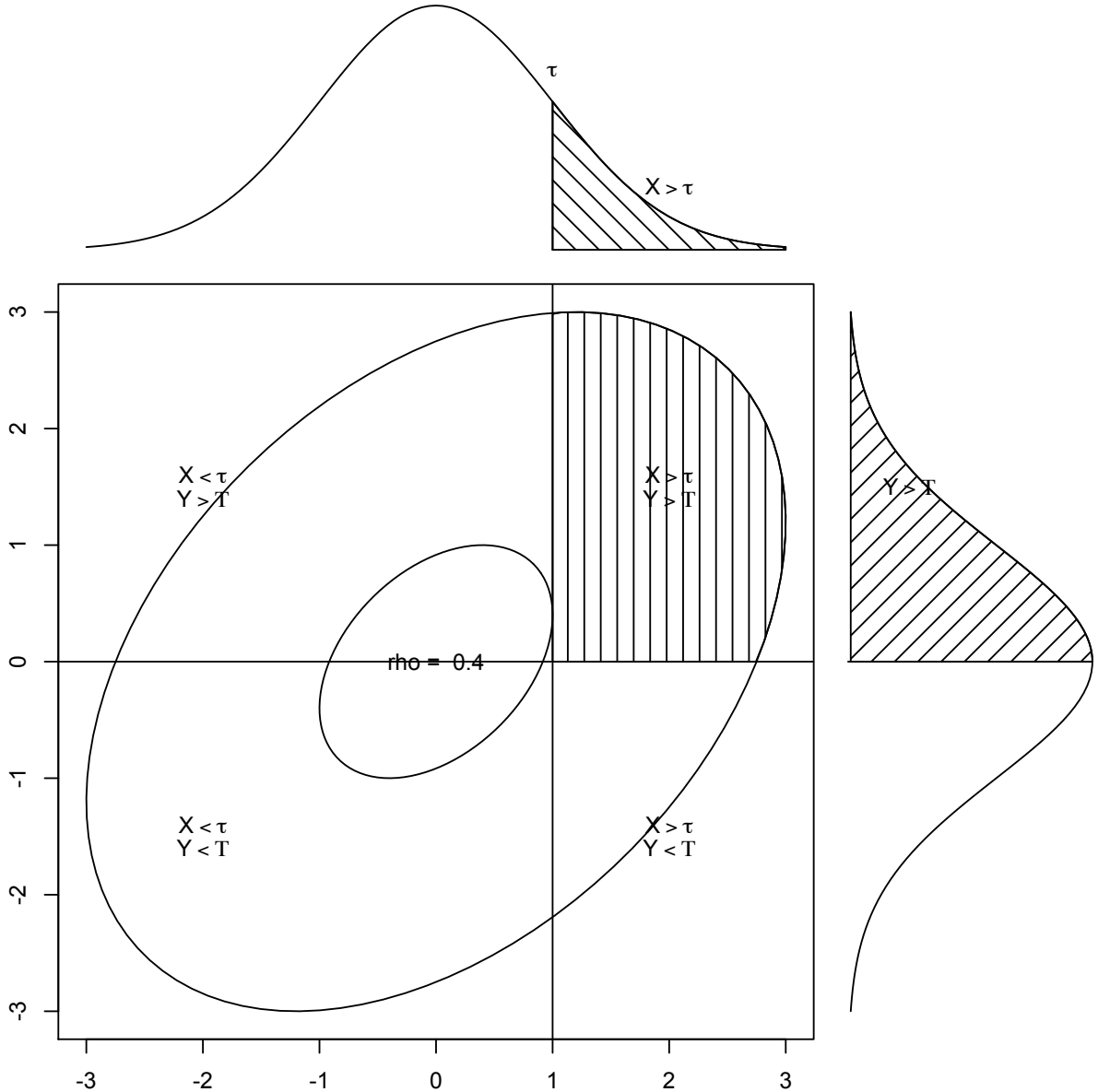
```
> sex <- c(2, 1041,0,6257)
> phi(sex)

[1] 0.04
```

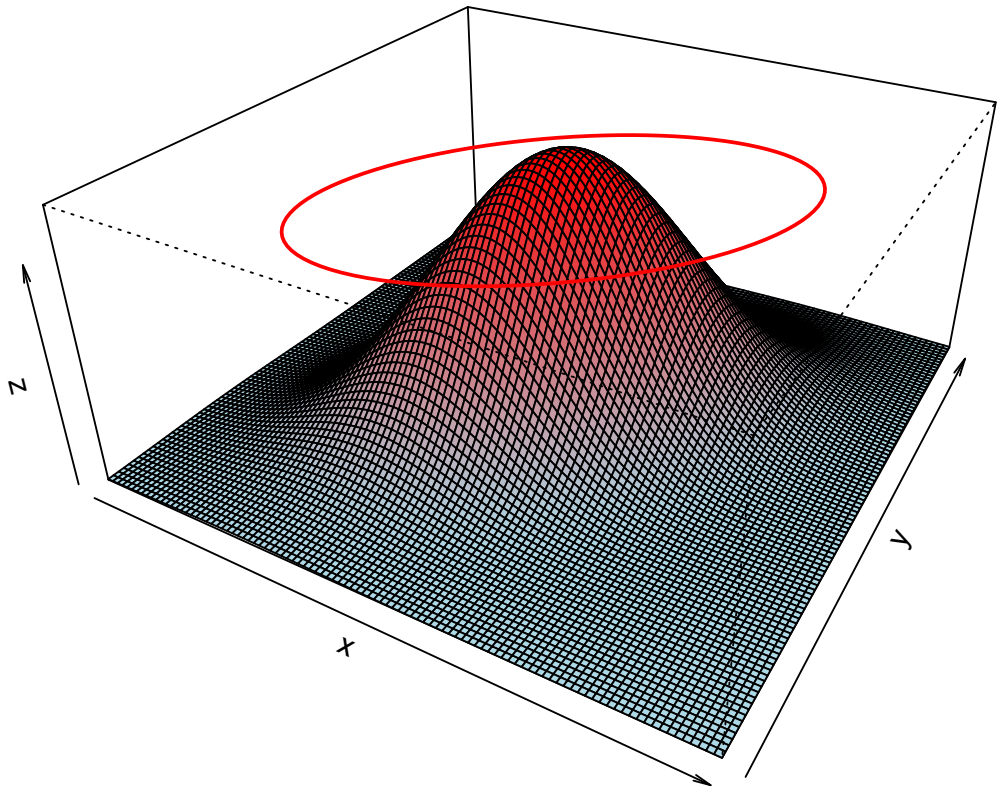
The biserial correlation estimates the latent correlation



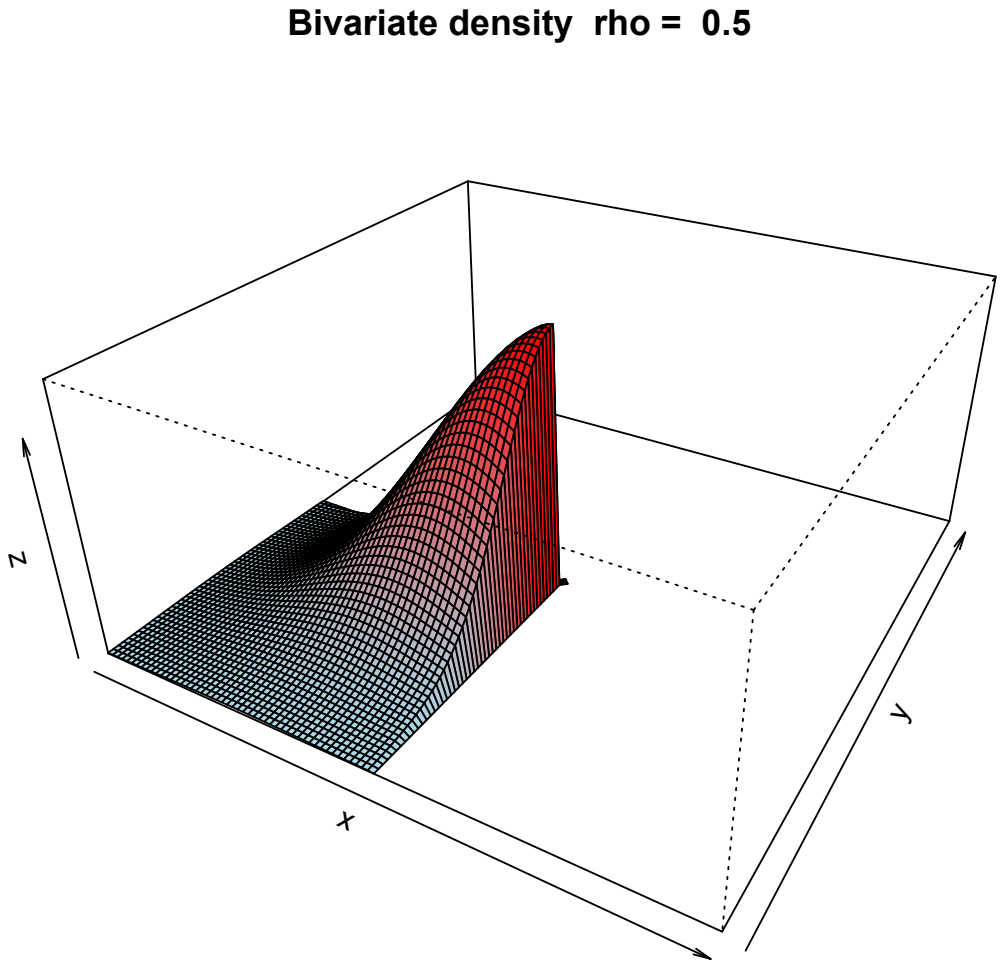
The tetrachoric correlation estimates the latent correlation



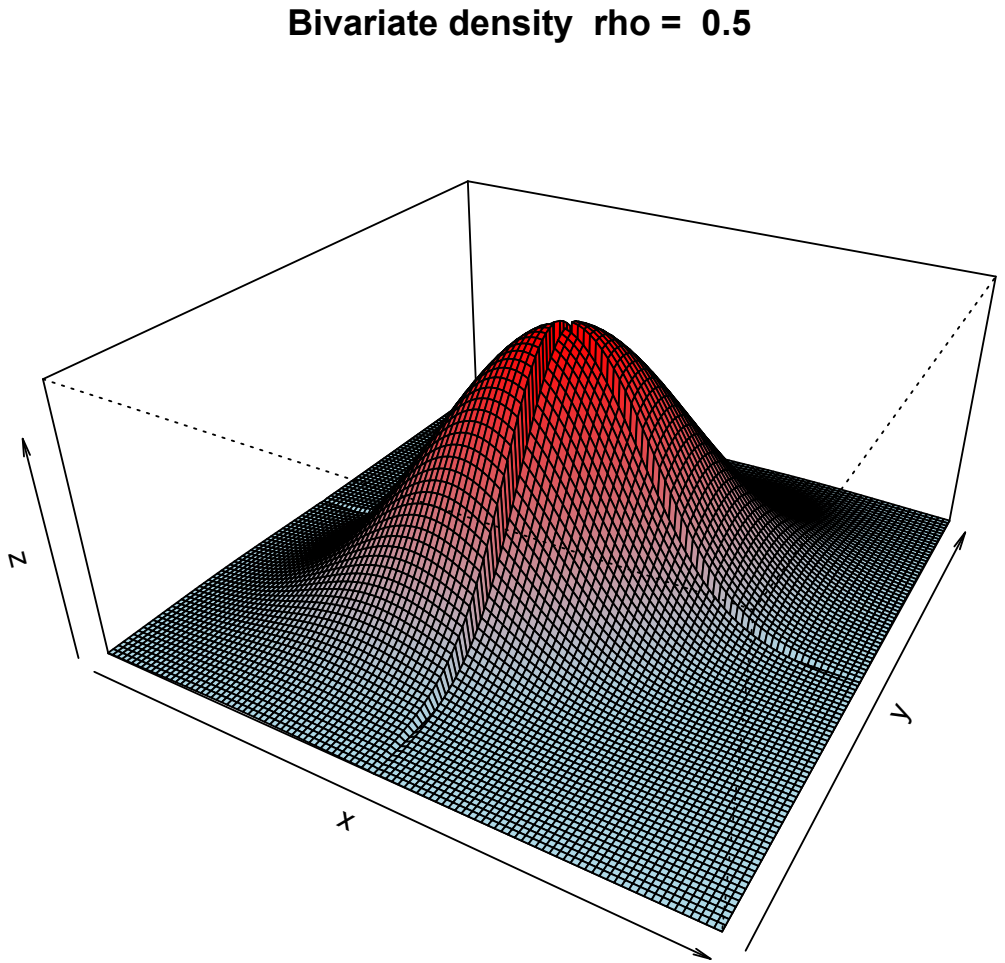
A bivariate normal correlation



Estimating a bivariate normal with 1 cell of a 2 x 2 table



Estimating a bivariate normal with 4 cells of a 2 x 2 table



Correlation size $\neq$ causal importance – tetrachoric correlation

Table : The relationship between sex and pregnancy (hypothetical data)

	Pregnant	Not Pregnant	Total
Intercourse	2	1,041	1,043
No intercourse	0	6,257	6,257
Total	2	7,298	7,300
Phi	.04	ρ_{tet}	.95

```
> sex <- c(2, 1041,0,6257)
> phi(sex)
[1] 0.04
> tetrachoric(sex,correct=FALSE)
Call: tetrachoric(x = sex, correct = FALSE)
tetrachoric correlation
[1] 0.95

 with tau of
[1] -3.5 -1.1
```

Pearson r versus tetrachoric correlation on dichotomous ability data

```
> tet <- tetrachoric(ability)  #find the tetrachoric correlations
Loading required package: mvtnorm
Loading required package: parallel
> per <- lowerCor(ability)   #find the normal Pearson correlations
> per.tet <- lowerUpper(tet$rho,per) #combine them with tetrachorics below diagonal
> per.tet.diff <- lowerUpper(tet$rho,per,diff=TRUE) #show the differences
```

```
> round(per.tet[1:8,1:8],2)  #round to 2 decimal places
          reason.4 reason.16 reason.17 reason.19 letter.7 letter.33 letter.34 letter.58
reason.4        NA      0.28      0.40      0.30     0.28      0.23      0.29      0.29
reason.16     0.45        NA      0.32      0.25     0.27      0.20      0.26      0.21
reason.17     0.61      0.51        NA      0.34     0.29      0.26      0.29      0.29
reason.19     0.46      0.40      0.53        NA     0.25      0.25      0.27      0.25
letter.7      0.45      0.43      0.47      0.40       NA      0.34      0.40      0.33
letter.33     0.37      0.32      0.42      0.39     0.52        NA      0.37      0.28
letter.34     0.46      0.41      0.47      0.43     0.60      0.56        NA      0.32
letter.58     0.47      0.35      0.48      0.40     0.51      0.43      0.50        NA
> round(per.tet.diff[1:8,1:8],2)  #show the differences
          reason.4 reason.16 reason.17 reason.19 letter.7 letter.33 letter.34 letter.58
reason.4        NA      0.17      0.21      0.17     0.16      0.14      0.17      0.18
reason.16     0.45        NA      0.19      0.15     0.16      0.13      0.16      0.14
reason.17     0.61      0.51        NA      0.19     0.18      0.16      0.18      0.19
reason.19     0.46      0.40      0.53        NA     0.14      0.14      0.15      0.15
letter.7      0.45      0.43      0.47      0.40       NA      0.18      0.20      0.18
letter.33     0.37      0.32      0.42      0.39     0.52        NA      0.19      0.15
letter.34     0.46      0.41      0.47      0.43     0.60      0.56        NA      0.18
letter.58     0.47      0.35      0.48      0.40     0.51      0.43      0.50        NA
```

Pearson r versus polychoric correlation on 6 alternative BFI data

```
> poly <- polychoric(bfi[1:10]) #polychorics on just the first 10 variables
> pearson <- cor(bfi[1:10],use="pairwise") #Pearson on the first 10
> poly.pear <- lowerUpper(poly$rho,pearson)
> poly.pear.diff <- lowerUpper(poly$rho,pearson,diff=TRUE) #show differences
> poly.pear
> round(poly.pear,2)
      A1    A2    A3    A4    A5    C1    C2    C3    C4    C5
A1    NA -0.34 -0.27 -0.15 -0.18  0.03  0.02 -0.02  0.13  0.05
A2 -0.41    NA  0.49  0.34  0.39  0.09  0.14  0.19 -0.15 -0.12
A3 -0.32  0.56    NA  0.36  0.50  0.10  0.14  0.13 -0.12 -0.16
A4 -0.18  0.39  0.41    NA  0.31  0.09  0.23  0.13 -0.15 -0.24
A5 -0.23  0.45  0.57  0.36    NA  0.12  0.11  0.13 -0.13 -0.17
C1  0.00  0.12  0.12  0.11  0.16    NA  0.43  0.31 -0.34 -0.25
C2  0.01  0.16  0.16  0.27  0.14  0.48    NA  0.36 -0.38 -0.30
C3 -0.02  0.23  0.16  0.17  0.15  0.34  0.40    NA -0.34 -0.34
C4  0.15 -0.19 -0.16 -0.20 -0.17 -0.40 -0.43 -0.38    NA  0.48
C5  0.06 -0.16 -0.19 -0.28 -0.20 -0.29 -0.33 -0.38  0.53    NA
> round(poly.pear.diff,2)
      A1    A2    A3    A4    A5    C1    C2    C3    C4    C5
A1    NA -0.07 -0.06 -0.03 -0.05 -0.02 -0.01  0.00  0.02  0.01
A2 -0.41    NA  0.07  0.05  0.06  0.02  0.02  0.03 -0.05 -0.03
A3 -0.32  0.56    NA  0.05  0.07  0.03  0.02  0.03 -0.04 -0.03
A4 -0.18  0.39  0.41    NA  0.05  0.02  0.04  0.04 -0.04 -0.04
A5 -0.23  0.45  0.57  0.36    NA  0.04  0.03  0.02 -0.04 -0.03
C1  0.00  0.12  0.12  0.11  0.16    NA  0.06  0.04 -0.06 -0.04
C2  0.01  0.16  0.16  0.27  0.14  0.48    NA  0.04 -0.05 -0.03
C3 -0.02  0.23  0.16  0.17  0.15  0.34  0.40    NA -0.04 -0.04
C4  0.15 -0.19 -0.16 -0.20 -0.17 -0.40 -0.43 -0.38    NA  0.05
C5  0.06 -0.16 -0.19 -0.28 -0.20 -0.29 -0.33 -0.38  0.53    NA
```

Spearman vs. Pearson on BFI data

```
> spear <- cor(bfi[1:10],use="pairwise",method="spearman")
> spear.pear <- lowerUpper(spear,pearson,diff=TRUE)
> round(spear.pear,2)
```

```
      A1    A2    A3    A4    A5    C1    C2    C3    C4    C5
A1    NA -0.03 -0.03 -0.01 -0.04 -0.05 -0.03 -0.02  0.02  0.01
A2 -0.37    NA  0.02  0.00  0.01  0.02  0.01  0.01 -0.03 -0.03
A3 -0.30  0.50    NA  0.00  0.03  0.02  0.01  0.02 -0.03 -0.02
A4 -0.16  0.34  0.36    NA  0.01  0.01  0.02  0.02 -0.03 -0.01
A5 -0.22  0.40  0.53  0.31    NA  0.02  0.02  0.01 -0.03 -0.02
C1 -0.02  0.11  0.12  0.10  0.15    NA  0.02  0.01 -0.04 -0.01
C2 -0.01  0.14  0.15  0.25  0.13  0.45    NA  0.01 -0.02  0.00
C3 -0.04  0.21  0.16  0.15  0.14  0.32  0.37    NA -0.01 -0.01
C4  0.15 -0.18 -0.16 -0.18 -0.16 -0.38 -0.40 -0.35    NA  0.01
C5  0.06 -0.15 -0.18 -0.26 -0.19 -0.26 -0.30 -0.35  0.49    NA
```

Comments on these alternative correlations

1. The assumption is that there was an underlying bivariate, normal distribution that was somehow artificially dichotomized.
2. But some things are in fact dichotomous, not normally distributed
   - Alive/Dead
   - Vacinated/Not vacinated
3. polychoric and tetrachoric correlations are found by iteratively fitting bivariate normal distributions with varying correlations until the best fit for a n x n table is found.
4. This is done using the `tetrachoric` or `polychoric` functions. They are not fast! (In comparison to Pearson r).

Cautions about correlations–The Anscombe data set

Consider the following 8 variables

	var	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosi
x1	1	11	9.0	3.32	9.00	9.00	4.45	4.00	14.00	10.00	0.00	-1.2
x2	2	11	9.0	3.32	9.00	9.00	4.45	4.00	14.00	10.00	0.00	-1.2
x3	3	11	9.0	3.32	9.00	9.00	4.45	4.00	14.00	10.00	0.00	-1.2
x4	4	11	9.0	3.32	8.00	8.00	0.00	8.00	19.00	11.00	2.47	11.0
y1	5	11	7.5	2.03	7.58	7.49	1.82	4.26	10.84	6.58	-0.05	-0.5
y2	6	11	7.5	2.03	8.14	7.79	1.47	3.10	9.26	6.16	-0.98	0.8
y3	7	11	7.5	2.03	7.11	7.15	1.53	5.39	12.74	7.35	1.38	4.3
y4	8	11	7.5	2.03	7.04	7.20	1.90	5.25	12.50	7.25	1.12	3.1

Cautions, Anscombe continued

With regressions of

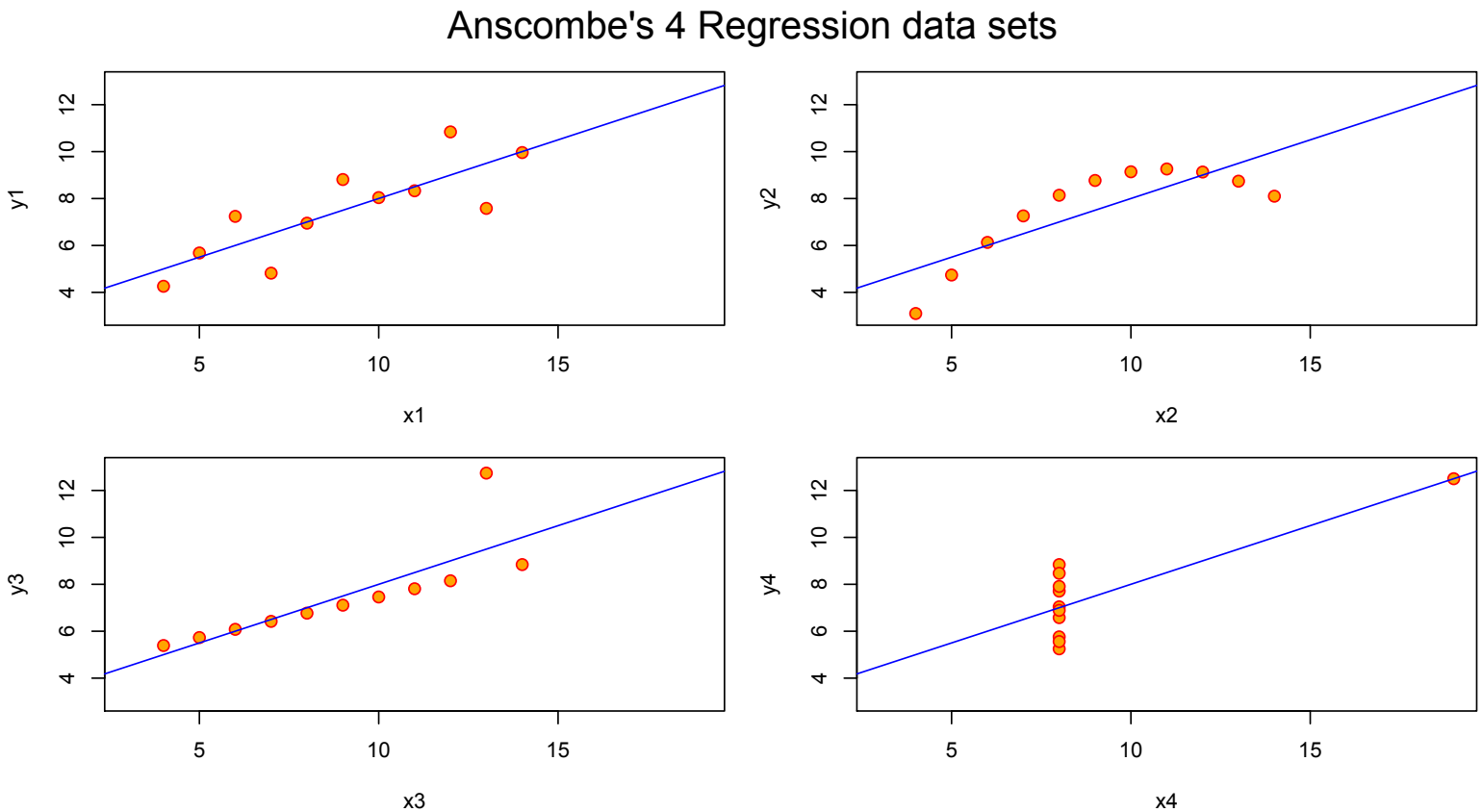
```
             Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 3.0000909  1.1247468 2.667348 0.025734051
x1          0.5000909  0.1179055 4.241455 0.002169629

[[2]]
             Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 3.000909  1.1253024 2.666758 0.025758941
x2          0.500000  0.1179637 4.238590 0.002178816

[[3]]
             Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 3.0024545  1.1244812 2.670080 0.025619109
x3          0.4997273  0.1178777 4.239372 0.002176305

[[4]]
             Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 3.0017273  1.1239211 2.670763 0.025590425
x4          0.4999091  0.1178189 4.243028 0.002164602
```

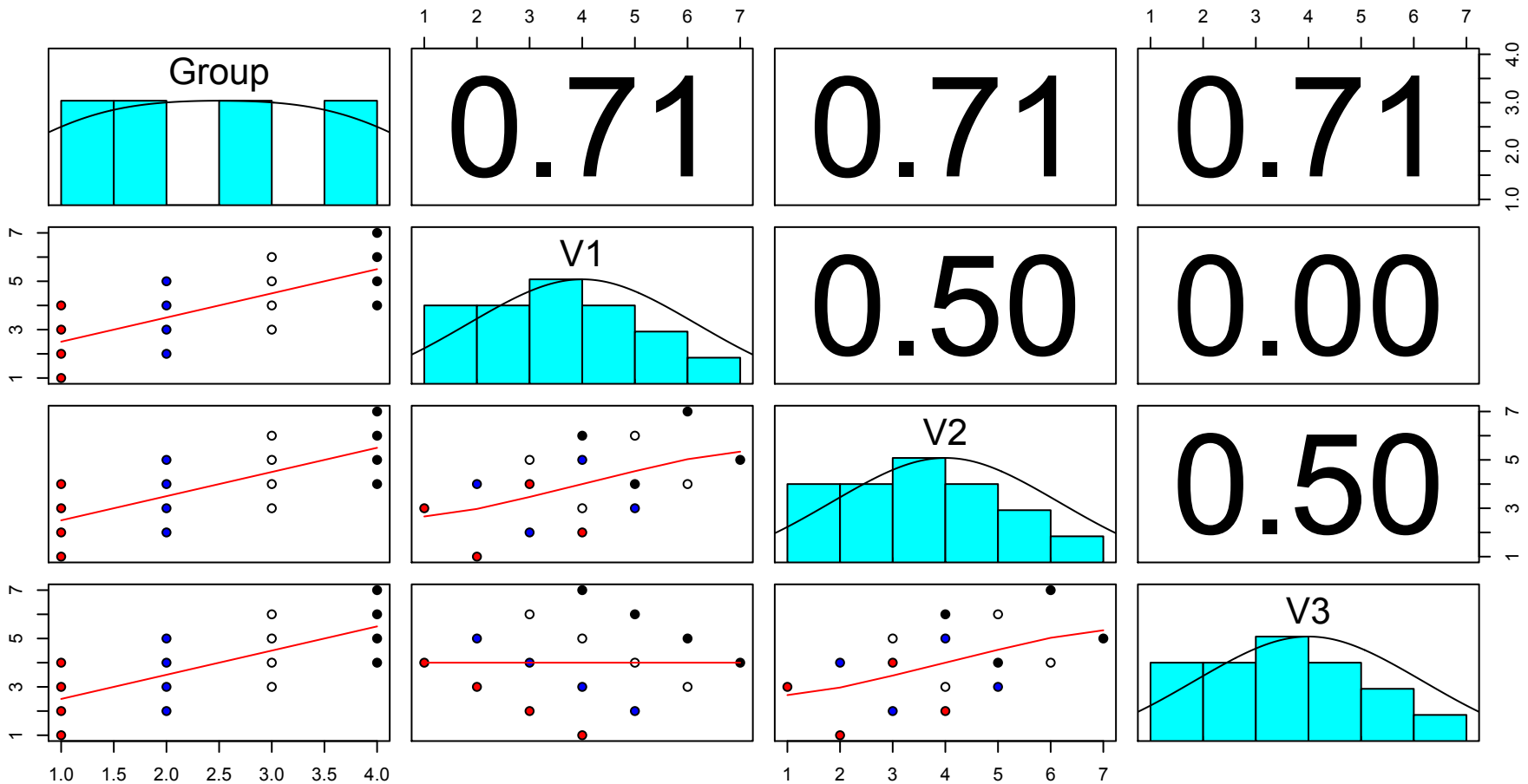
Cautions about correlations: Anscombe data set



Further cautions about correlations–the problem of levels

1. Correlations taken at one level of analysis can be unrelated to those at another level
2. $r_{xy} = \eta_{x_{wg}} * \eta_{y_{wg}} * r_{xy_{wg}} + \eta_{x_{bg}} * \eta_{y_{bg}} * r_{xy_{bg}}$
3. Where η is the correlation of the data with the within group values, or the group means.
4. The within group and between group correlations can even be of different sign!
5. The `withinBetween` data set is an example of this problem.
6. The `statsBy` function will find the within and between group correlations for this kind of multi-level design.

Cautions about correlations: Within versus between groups



Bias, or just Simpson's Paradox?

Table : Hypothetical Admissions data showing 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

Calculate the ϕ and tetrachoric correlations

```
> admit <- c(40,10,10,40)
> phi(admit)

[1] 0.6

> phi2poly(.6,.5,.5)

[1] 0.8090178

> tetrachoric(admit)

Call: tetrachoric(x = admit)
tetrachoric correlation
[1] 0.81

 with tau of
[1] 0 0
```

1. Input the four cell counts
2. Find the ϕ coefficient
3. Convert this to a tetrachoric correlation by specifying the marginals
4. Or, just call tetrachoric with these cell entries

Sex discrimination by department shows opposite effect

Table : Hypothetical Admissions data showing sex discrimination

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

Table : Males: unselective

	Admit	Reject	Total
Male	40	5	45
Female	5	0	5
Total	45	5	50
ϕ	-.11	ρ	-.95

Table : Females: selective

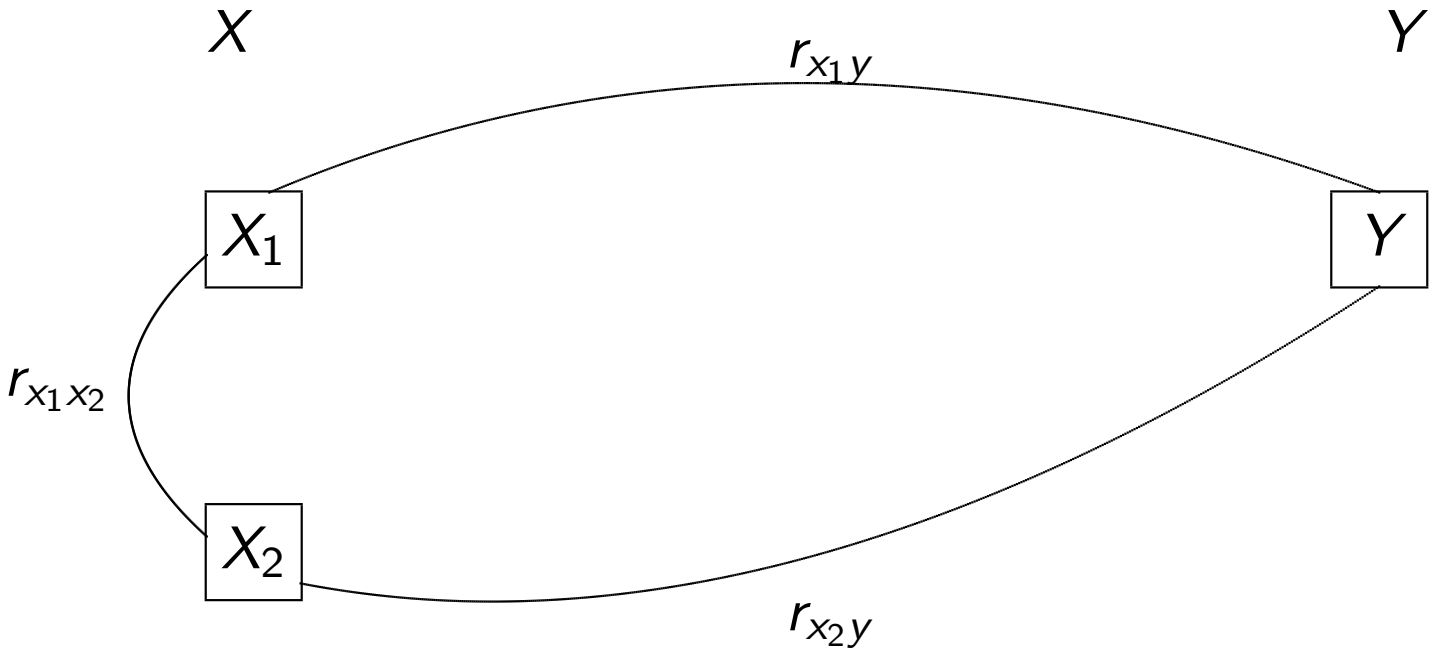
	Admit	Reject	Total
Male	0	5	5
Female	5	40	45
Total	5	45	50
ϕ	-.11	ρ	-.95

The ubiquitous correlation coefficient

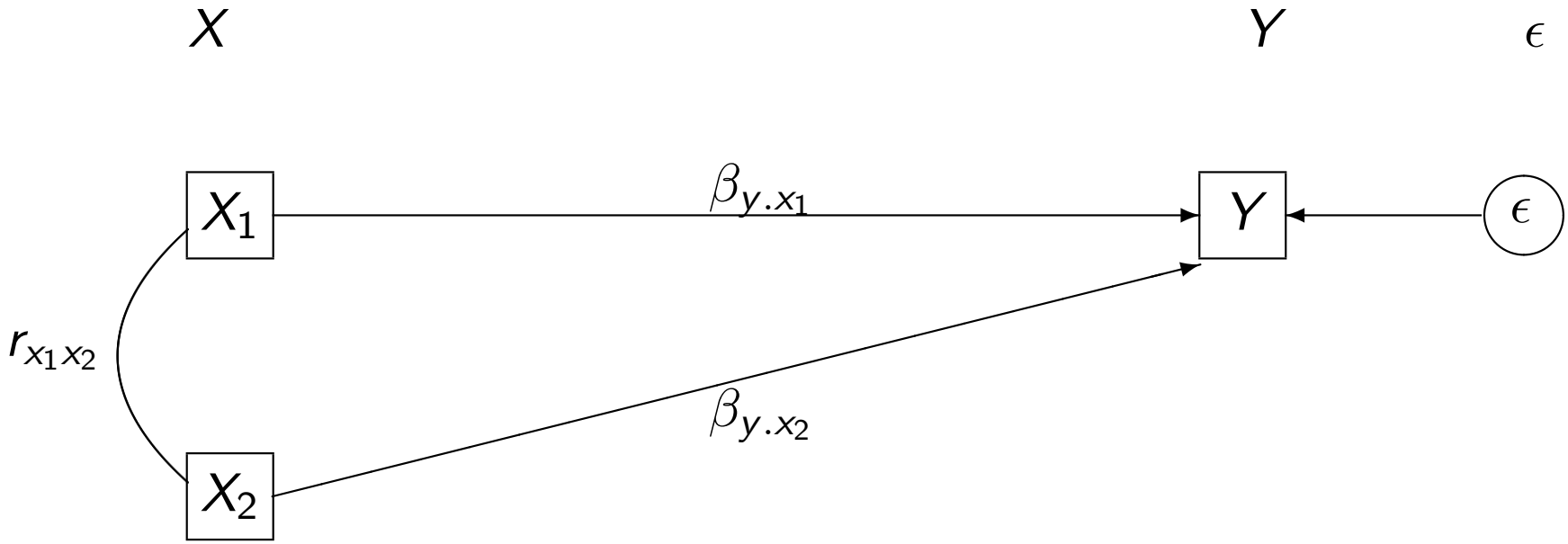
Table : Alternative Estimates of effect size. Using the correlation as a scale free estimate of effect size allows for combining experimental and correlational data in a metric that is directly interpretable as the effect of a standardized unit change in x leads to r change in standardized y.

Statistic	Estimate	r equivalent	as a function of r
Pearson correlation	$r_{xy} = \frac{C_{xy}}{\sigma_x \sigma_y}$	r_{xy}	
Regression	$b_{y.x} = \frac{C_{xy}}{\sigma_x^2}$	$r = b_{y.x} \frac{\sigma_y}{\sigma_x}$	$b_{y.x} = r \frac{\sigma_x}{\sigma_y}$
Cohen's d	$d = \frac{X_1 - X_2}{\sigma_x}$	$r = \frac{d}{\sqrt{d^2+4}}$	$d = \frac{2r}{\sqrt{1-r^2}}$
Hedge's g	$g = \frac{X_1 - X_2}{s_x}$	$r = \frac{g}{\sqrt{g^2+4(df/N)}}$	$g = \frac{2r\sqrt{df/N}}{\sqrt{1-r^2}}$
t - test	$t = \frac{d\sqrt{df}}{2}$	$r = \sqrt{t^2/(t^2+df)}$	$t = \sqrt{\frac{r^2 df}{1-r^2}}$
F-test	$F = \frac{d^2 df}{4}$	$r = \sqrt{F/(F+df)}$	$F = \frac{r^2 df}{1-r^2}$
Chi Square		$r = \sqrt{\chi^2/n}$	$\chi^2 = r^2 n$
Odds ratio	$d = \frac{ln(OR)}{1.81}$	$r = \frac{ln(OR)}{1.81\sqrt{(ln(OR)/1.81)^2+4}}$	$ln(OR) = \frac{3.62r}{\sqrt{1-r^2}}$
$r_{equivalent}$	r with probability p	$r = r_{equivalent}$	

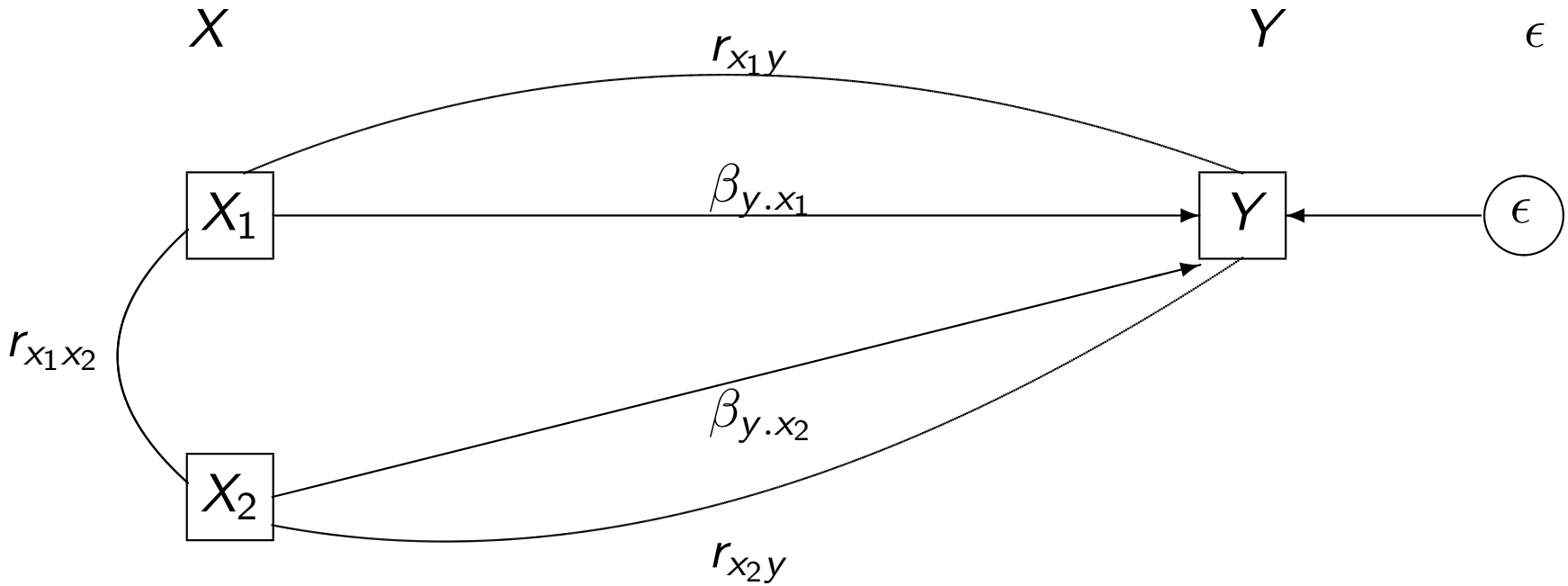
Multiple correlations



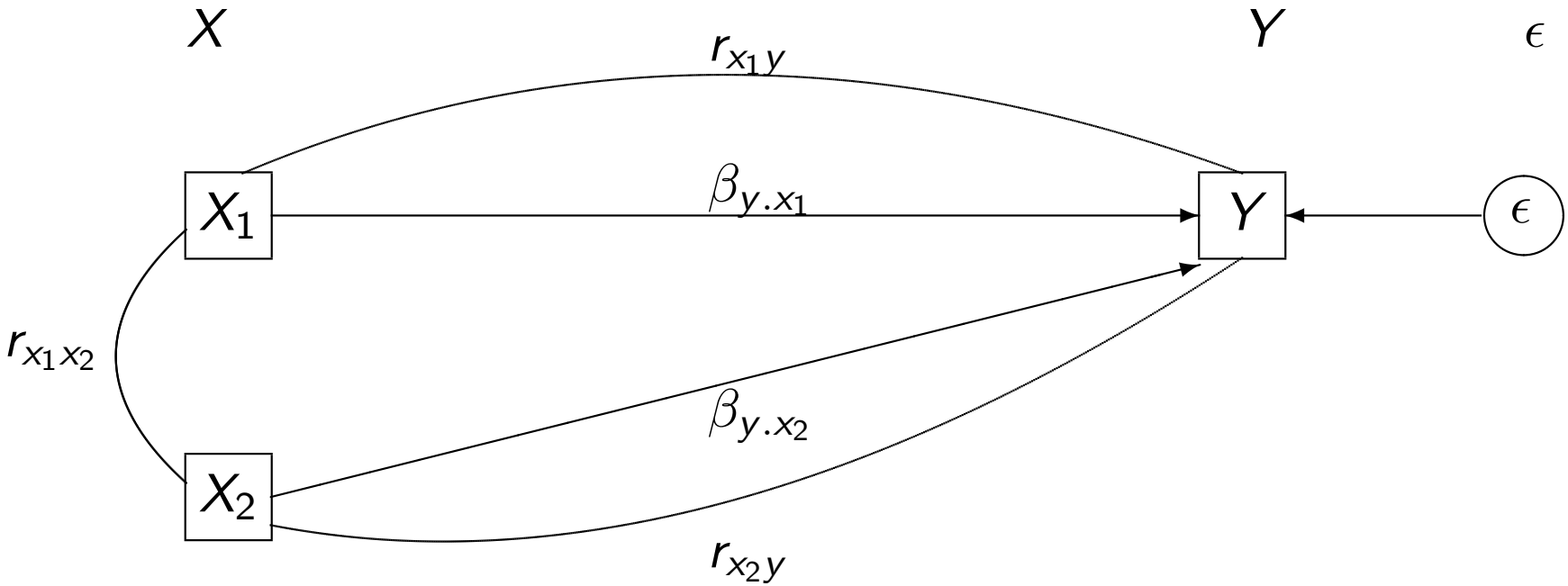
Multiple Regression



Multiple Regression: decomposing correlations



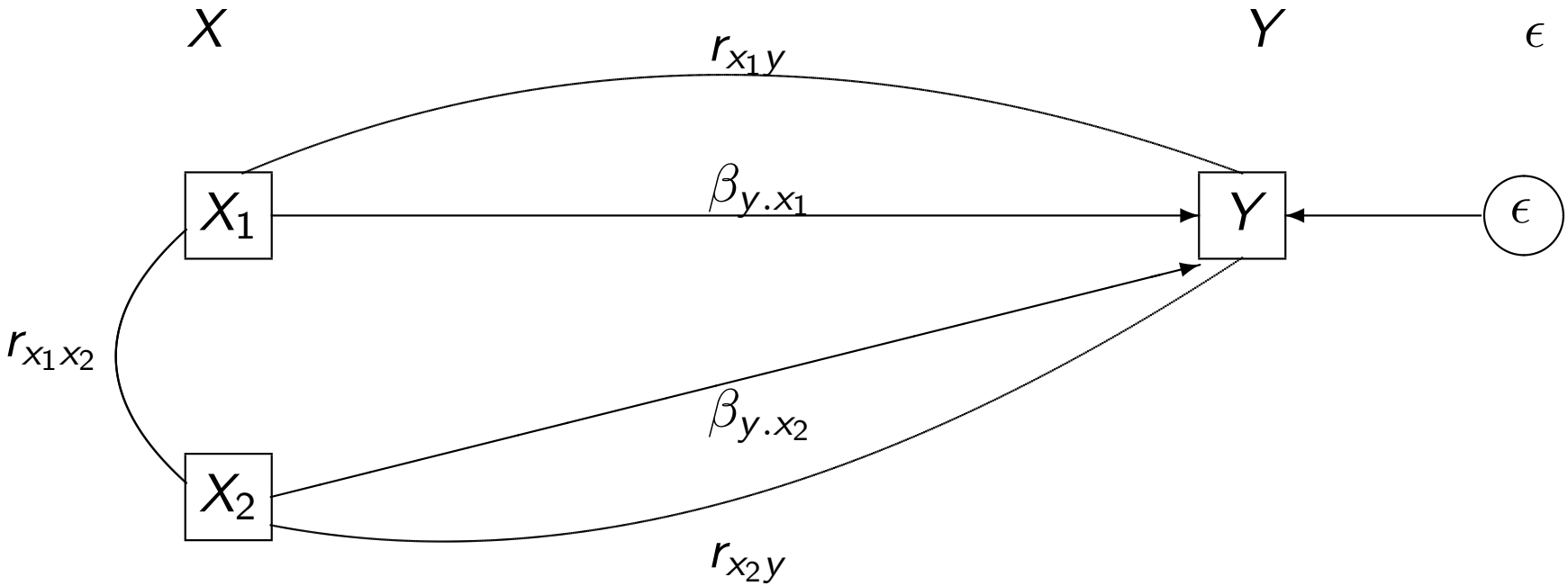
Multiple Regression: decomposing correlations



$$r_{x_1 y} = \overbrace{\beta_{y.x_1}}^{direct} + \overbrace{r_{x_1 x_2}\beta_{y.x_2}}^{indirect}$$

$$r_{x_2 y} = \underbrace{\beta_{y.x_2}}_{direct} + \underbrace{r_{x_1 x_2}\beta_{y.x_1}}_{indirect}$$

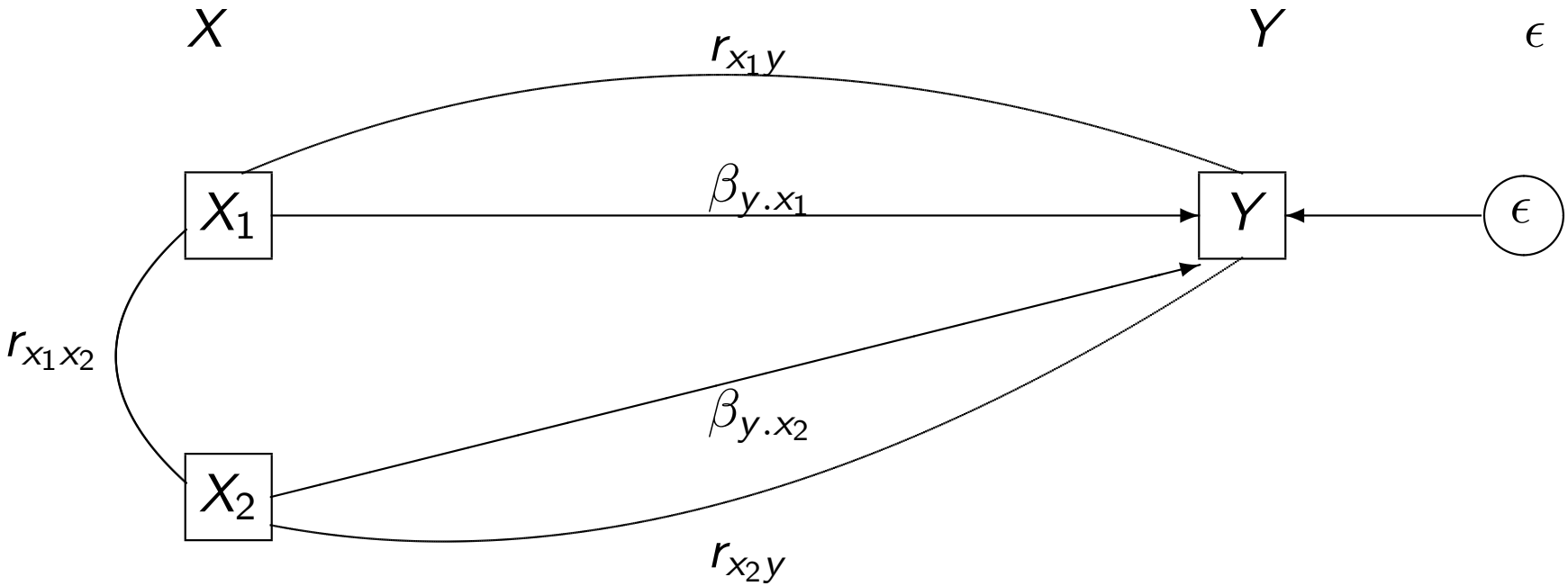
Multiple Regression: decomposing correlations



$$r_{x_1y} = \overbrace{\beta_{y.x_1}}^{direct} + \overbrace{r_{x_1x_2}\beta_{y.x_2}}^{indirect} \qquad \beta_{y.x_1} = \frac{r_{x_1y} - r_{x_1x_2}r_{x_2y}}{1 - r^2_{x_1x_2}}$$

$$r_{x_2y} = \underbrace{\beta_{y.x_2}}_{direct} + \underbrace{r_{x_1x_2}\beta_{y.x_1}}_{indirect} \qquad \beta_{y.x_2} = \frac{r_{x_2y} - r_{x_1x_2}r_{x_1y}}{1 - r^2_{x_1x_2}}$$

Multiple Regression: decomposing correlations



$$r_{x_1y} = \overbrace{\beta_{y.x_1}}^{direct} + \overbrace{r_{x_1x_2}\beta_{y.x_2}}^{indirect}$$

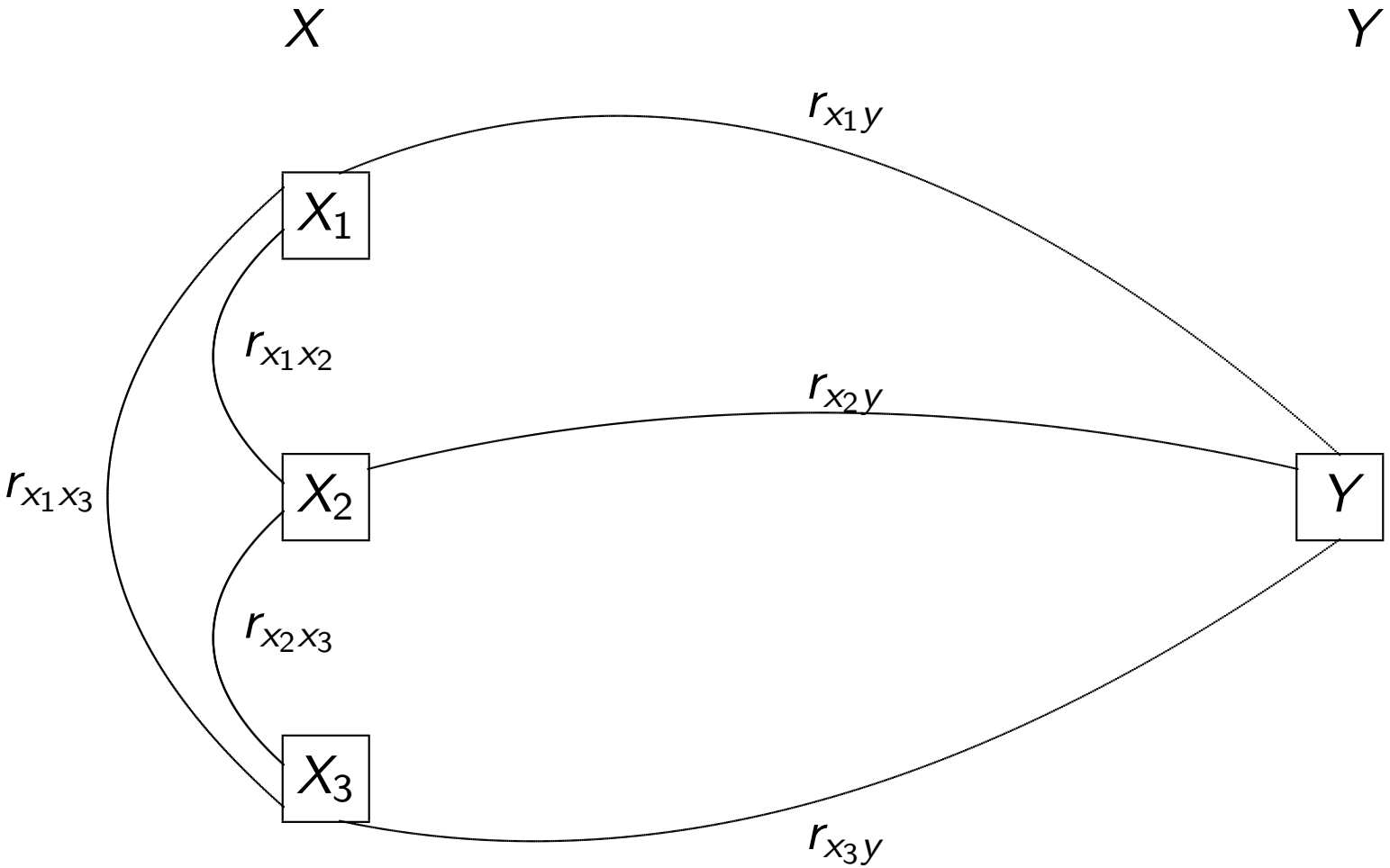
$$r_{x_2y} = \underbrace{\beta_{y.x_2}}_{direct} + \underbrace{r_{x_1x_2}\beta_{y.x_1}}_{indirect}$$

$$\beta_{y.x_1} = \frac{r_{x_1y} - r_{x_1x_2}r_{x_2y}}{1 - r^2_{x_1x_2}}$$

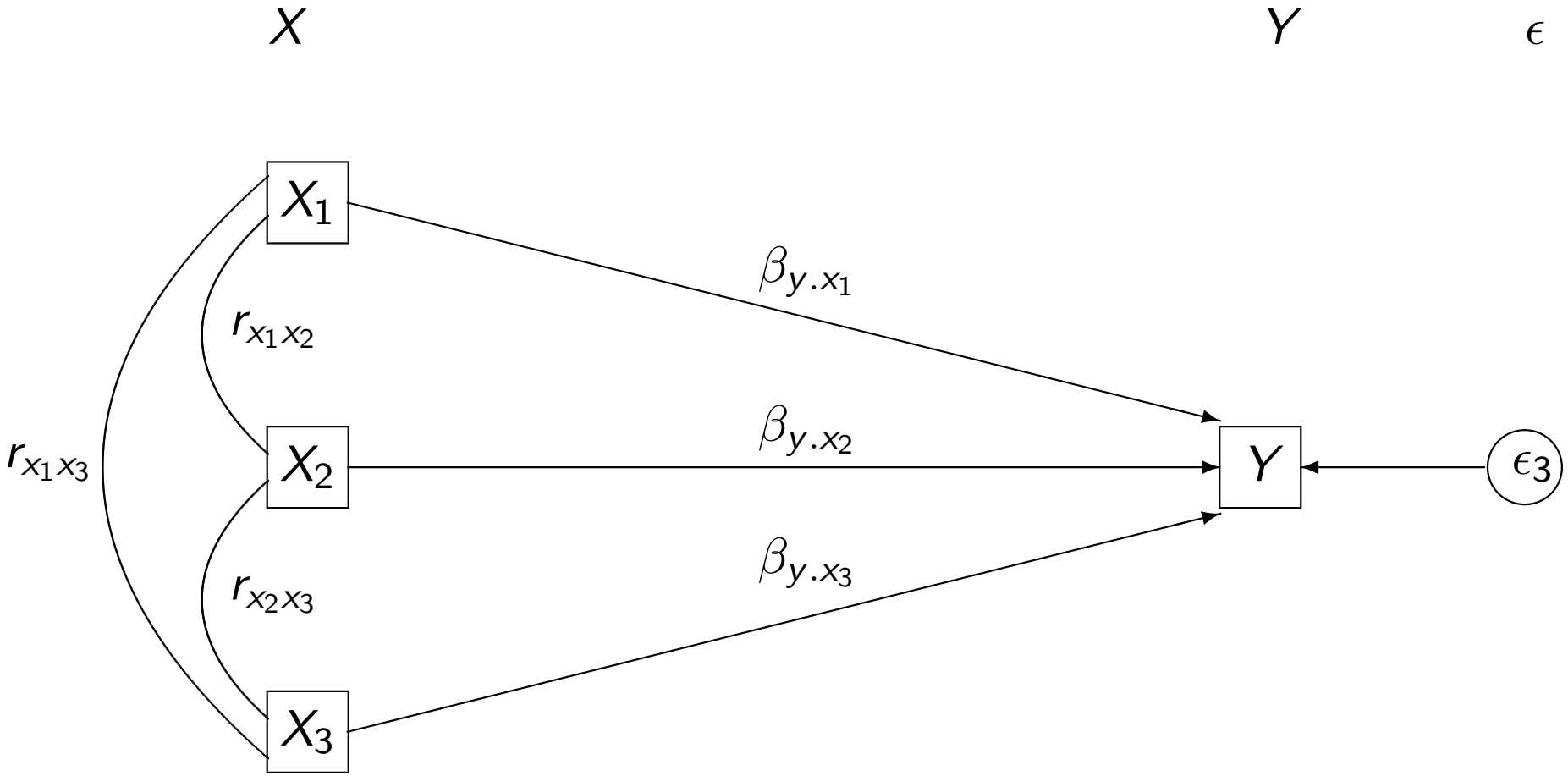
$$\beta_{y.x_2} = \frac{r_{x_2y} - r_{x_1x_2}r_{x_1y}}{1 - r^2_{x_1x_2}}$$

$$R^2 = r_{x_1y}\beta_{y.x_1} + r_{x_2y}\beta_{y.x_2}$$

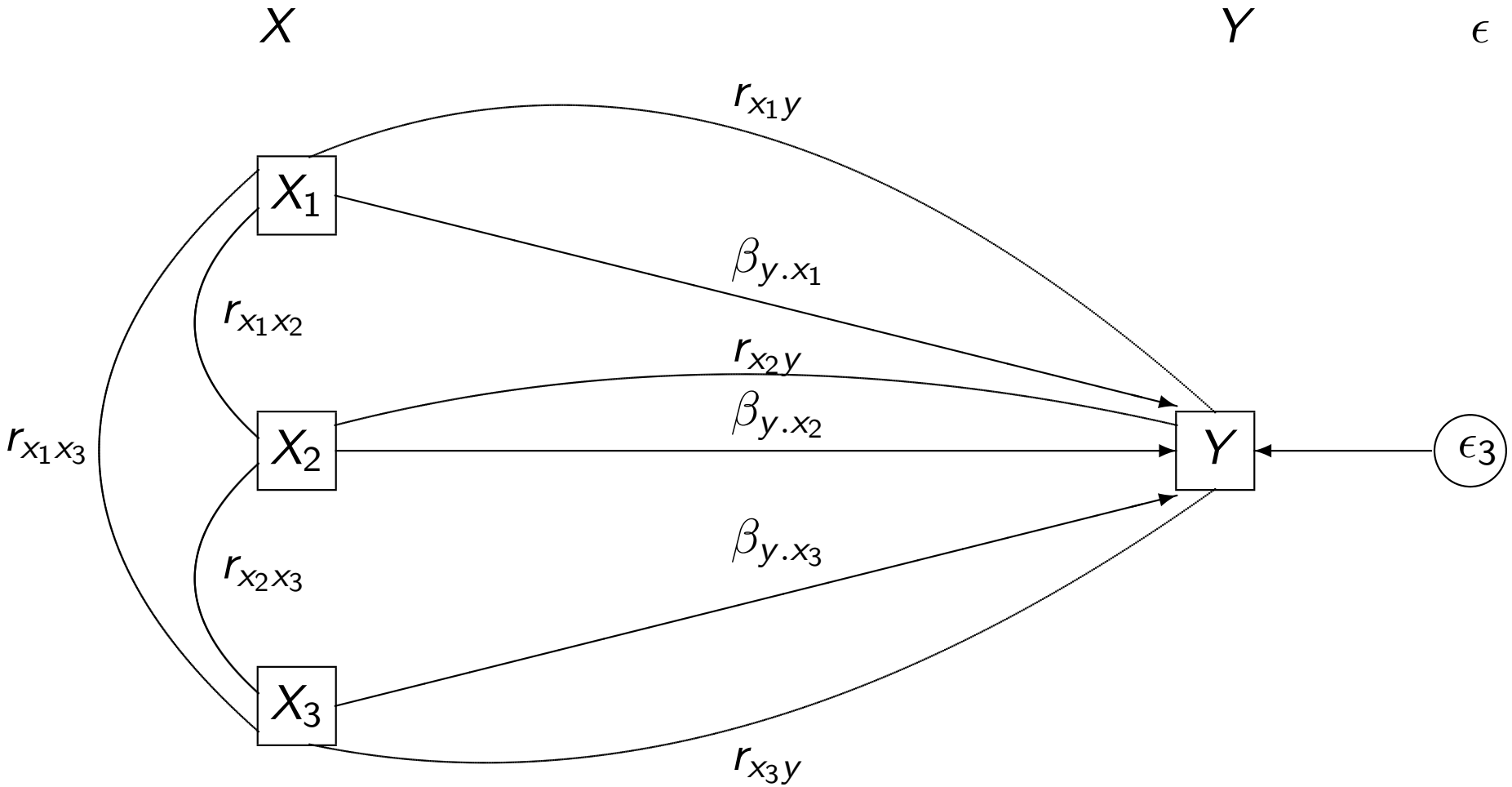
What happens with 3 predictors? The correlations



What happens with 3 predictors? β weights



What happens with 3 predictors?



$$r_{x_1y} = \underbrace{\beta_{y.x_1}}_{\textit{direct}} + \underbrace{r_{x_1x_2}\beta_{y.x_1} + r_{x_1x_3}\beta_{y.x_3}}_{\textit{indirect}} \qquad r_{x_2y} = \ldots \qquad r_{x_3y} = \ldots$$

The math gets tedious

Multiple regression and linear algebra

- Multiple regression requires solving multiple, simultaneous equations to estimate the direct and indirect effects.
  - Each equation is expressed as a $r_{x_i y}$ in terms of direct and indirect effects.
  - Direct effect is $\beta_{y.x_i}$
  - Indirect effect is $\sum_{j \neq i} beta_{y.x_j} r_{x_j y}$
- How to solve these equations?
- Tediously, or just use linear algebra.