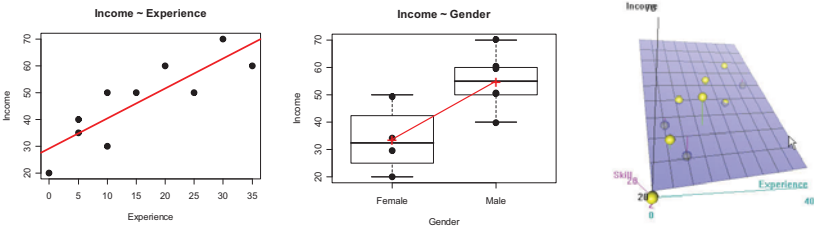
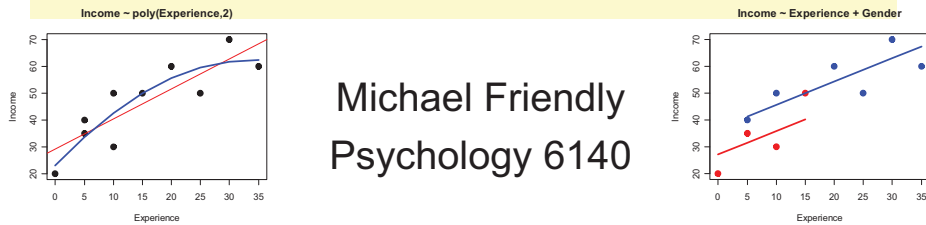


# Why study multivariate data analysis?

- Multivariate data more common in research
- GLM approach: ANOVA, regression, etc. within a common framework: linear models
 
$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} + \varepsilon_i$$
- In matrix form (  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$  ), GLM extends to MANOVA, MMRreg, etc.
- Idea of linear combinations extends readily to other methods: PCA, discriminant analysis, etc.
- Graphical methods, geometry → Insight



## Multivariate Data Analysis: Overview



Michael Friendly  
Psychology 6140

## Sample problem: workers' data

	Y	X1	X2	X3
Name	Income	Experience	Skill	Gender
1	Abby	20	0	2 Female
2	Betty	35	5	5 Female
3	Charles	40	5	8 Male
4	Doreen	30	10	6 Female
5	Ethan	50	10	10 Male
6	Francie	50	15	7 Female
7	Georges	60	20	12 Male
8	Harry	50	25	10 Male
9	Isaac	70	30	15 Male
10	Juan	60	35	13 Male

In truly multivariate data, we may have several outcomes:

- Income
- Job satisfaction
- Manager ratings
- etc.

## Linear models: Regression

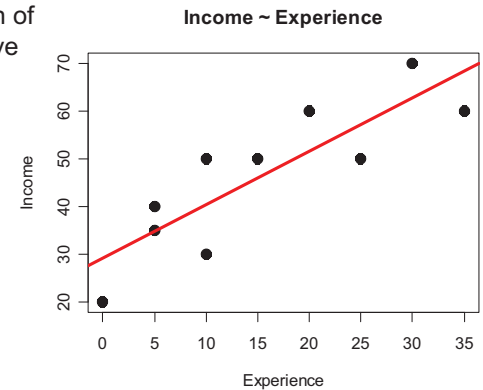
**Regression:** understanding the relation of quantitative predictor(s) on a quantitative outcome.

Model:  $E(y | x) = \beta_0 + \beta_1 x$   
e.g., Income = 29 + 1.12 Experience

Parameters:

$\beta_0 = 29 =$  Income at 0 years

$\beta_1 = 1.12 =$  Increase / year =  $\frac{\Delta y}{\Delta x}$



# Linear models: Regression

Regression: a “linear model” need only be **linear in the parameters**. It can have terms like  $x^2$ ,  $\log(x)$ , etc.

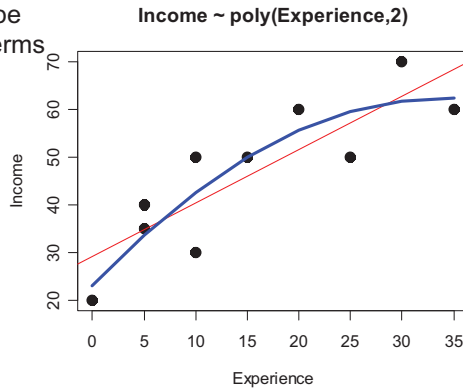
Model:  $E(y | x) = \beta_0 + \beta_1 x + \beta_2 x^2$   
 e.g,  $\text{Income} = 23 + 2.3 \text{Exp} - 0.33 \text{Exp}^2$

Parameters:

$\beta_0 = 23 = \text{Income at 0 years}$

$\beta_1 = 2.3 = \text{Slope at 0 years}$

$\beta_2 = -0.33 = \text{Decrease in slope/year}$



# Linear models: Multiple regression

Regression models can have **any number** of linear predictors

Model:  $E(y | x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

e.g,  $\text{Income} = 14.8 + 0.11 \text{Exper} + 3.4 \text{Skill}$

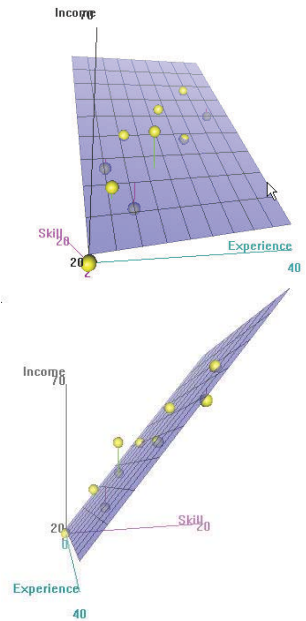
Parameters:

$\beta_0 = 14.8 = \text{Income at 0 years, 0 skill}$

$\beta_1 = 0.11 = \Delta \text{Income} / \Delta \text{Experience} \mid \text{Skill}$

$\beta_2 = 3.4 = \Delta \text{Income} / \Delta \text{Skill} \mid \text{Experience}$

Control: The estimated effect for each predictor controls (adjusts) for all others in the model



# Linear models: ANOVA

**ANOVA:** How does mean of quantitative response vary with a discrete factor?

Model:  $E(Y) = \mu + \beta (G='Male')$

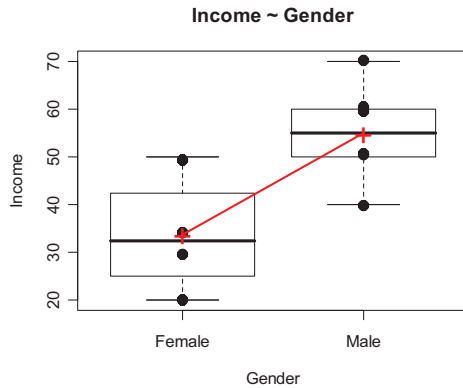
e.g.,  $\text{Income} = 33.75 + 21.25 (G='Male')$

$$(G = 'Male') \equiv \begin{pmatrix} 1 \\ 0 \end{pmatrix} \begin{matrix} M \\ F \end{matrix}$$

Parameters:

$\mu = 33.75 = \text{Female mean Income}$

$\beta = 21.25 = \text{Increment for Male}$



# Linear models: Regression + Anova

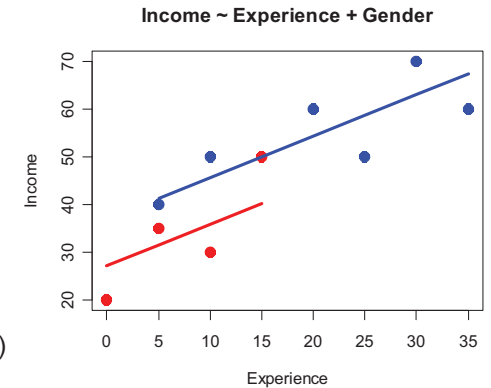
**ANCOVA:** Is there a difference in a factor, controlling for a quantitative predictor?

**Homogeneity of regression:** Are the regression lines for two or more groups the same? Are they parallel?

Model:  $E(Y) = \mu + \beta_1 X_1 + \beta_2 (G='Male')$

e.g.,

$\text{Inc} = 27.27 + 0.86 \text{Exp} + 9.73 (G='Male')$



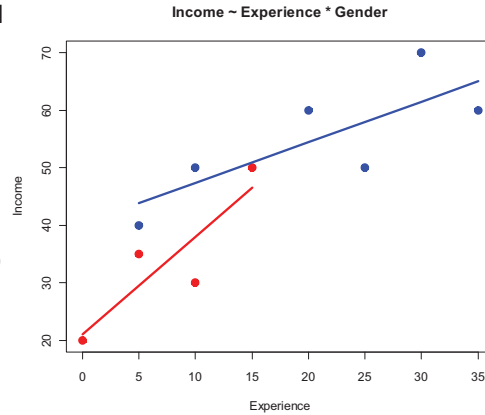
# Linear models: Regression + Anova

**Homogeneity of regression:** Test equal slopes by allowing a different slope for each group [X \* Group interaction]

Model:  $E(Y) = \mu + \beta_1 X_1 + \beta_2 (G='Male') + \beta_3 X_1 * (G='Male')$

e.g.,

$Inc = 21.0 + 1.70 Exp + 19.25 (G='Male') - 1.0 Exp * (G='Male')$



Thus, we have two separate models:

Females:  $Inc = 21.0 + 1.7 Exp$

Males:  $Inc = (21+19.25) + (1.7-1.0) Exp = 40.25 + 0.7 Exp$

# Linear models: Regression vs. ANOVA

	Regression	ANOVA
Dependent (response)	Quantitative	Quantitative
Independent (predictors)	Quantitative	Discrete factors
Concepts, statistics	Terms: $X_1, X_2$ Interactions: $X_1 * X_2$ Linear hypotheses $R^2$ , coefficients	Main effects: A, B Interactions: A*B Contrasts F stats, factor effects

# General Linear Model (GLM)

All of these are special cases of the **General Linear Model**:

Outcome = linear combination of predictors + residual

$$\begin{matrix} \text{Outcome} \\ \text{data} \end{matrix} = \underbrace{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}}_{\text{fitted (explained part)}} + \underbrace{\varepsilon_i}_{\text{residual (unexplained)}}$$

$\varepsilon_i \sim N(0, \sigma^2)$

where,

	Regression	ANOVA
$X$	Quantitative predictor (experience, skill)	Indicator (0/1) variables for group membership
$\beta$	Effect of predictor ( $\Delta y / \Delta x$ )	Diff between 0-group and 1-group

# General Linear Model (GLM)

They all become unified when cast in matrix terms:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{bmatrix} 1 & x_{11} & \dots \\ 1 & x_{21} & \dots \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

or,

$$y_{n \times 1} = X_{n \times (p+1)} \beta_{(p+1) \times 1} + \varepsilon_{n \times 1}$$

For all cases:

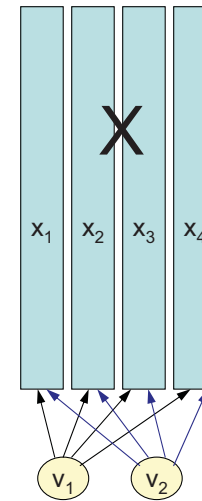
- parameter estimates, std. errors, etc. have the same form
- all hypothesis tests are special cases of  $H_0 : C \beta = 0$
- methods extend directly to: multivariate  $Y$ , non-normal errors, etc.

## Linear models & linear combinations

- All methods of multivariate statistics involve **linear combinations** of variables, with **weights** (coefficients) chosen to **optimize some criterion** (measure of fit)
- Methods differ according to:
  - 1 set of variables (PCA, FA) vs. 2+ sets (GLM, canonical correlation, discrim. analysis)
  - Nature of variables (2 sets):
    - Xs: discrete / continuous
    - Ys: discrete / continuous

13

## Linear combinations: 1 set of variables



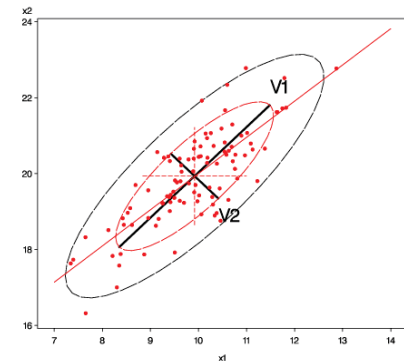
PCA: find weights to maximize variance of  $v_1, v_2, \dots$

$$v_1 = a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 x_4$$

$$v_2 = b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4$$

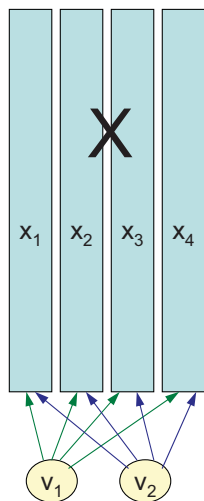
subject to: all  $v_i, v_j$  uncorrelated

PCA: Linear combinations to maximize variance



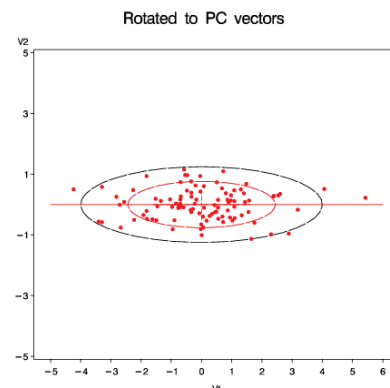
14

## Linear combinations: 1 set of variables



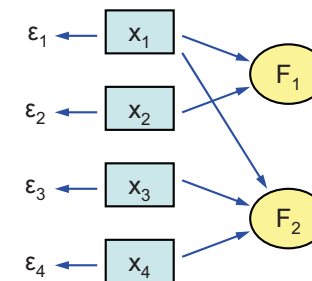
With  $p$  variables,  $p$  components account for 100% of variance, and correspond to a rotation of the variable space to uncorrelated components.

Goal in PCA is to account for most variance with  $k \ll p$  components.



15

## Factor analysis: Latent variables



FA: find weights for latent (unobserved) factors to account for correlations among observed variables

$$x_1 = \lambda_{11} F_1 + \lambda_{12} F_2 + \epsilon_1$$

$$x_2 = \lambda_{21} F_1 + \epsilon_2$$

$$x_3 = \lambda_{32} F_2 + \epsilon_3$$

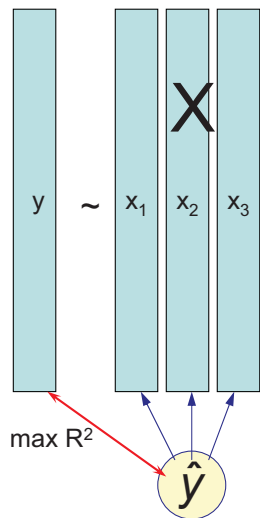
$$x_4 = \lambda_{42} F_2 + \epsilon_4$$

Differs from PCA in that error variance is taken into account.

FA can often give a simpler account with fewer factors or non-zero weights

16

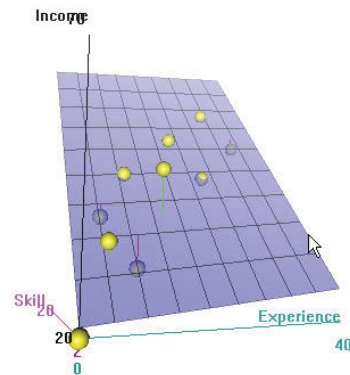
## Linear combinations: 2 sets of variables



Univariate response:

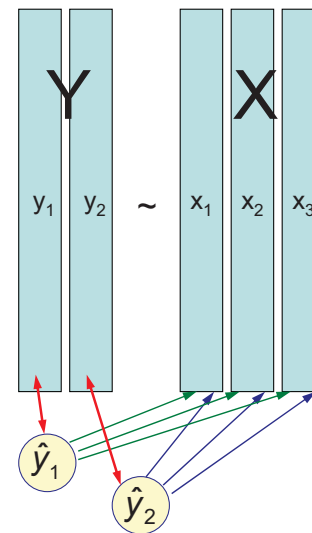
**MRA:** find weights to maximize correlation (R) between  $y$  and predicted  $y$ ,

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$



17

## 2 sets, multivariate response: MMRA



**Multivariate response: MMRA**

Multivariate MRA: find weights to maximize correlation between *each*  $y$  and predicted  $y$ ,

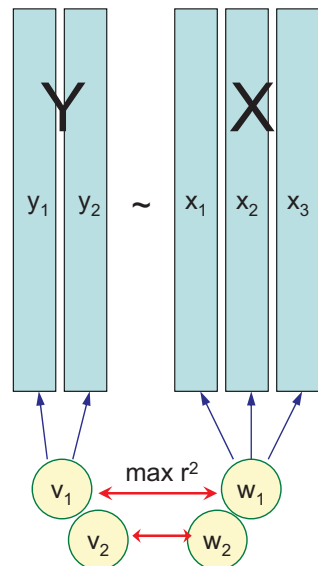
$$\hat{y}_1 = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

$$\hat{y}_2 = c_0 + c_1x_1 + c_2x_2 + c_3x_3$$

- Coefficients for each response are the same as in separate MRAs
- But: Multivariate tests take correlations among the  $y$ 's into account. Can be more powerful, by “pooling strength.”

18

## 2 sets, multivariate response: CanCorr



**Canonical correlation:**

Find linear combinations of the  $x$ 's that best predicts linear combination of the  $y$ 's

$$v_1 = a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 x_4$$

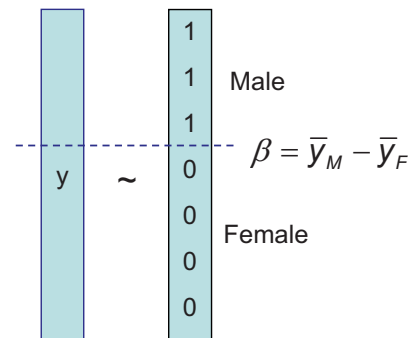
$$w_1 = b_1 y_1 + b_2 y_2 + b_3 y_3$$

- Choose weights to maximize  $r^2$  ( $v_1, w_1$ )
- Up to  $s = \min(p, q)$  additional pairs of canonical variables:  $(v_2, w_2), \dots (v_s, w_s)$
- All correlations between the  $Y$ s and  $X$ s are explained thru the correlation of each  $v_i$  with  $w_i$ .

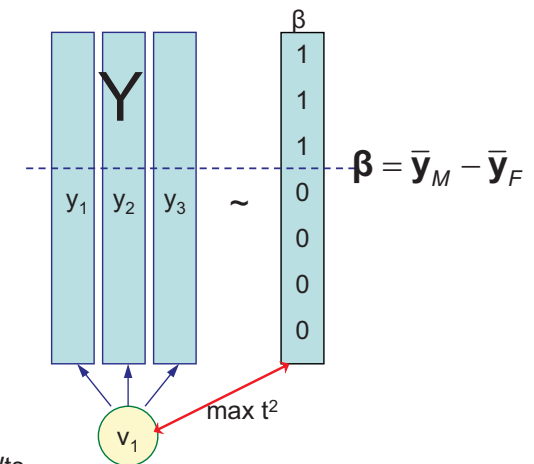
19

## Discrete predictors: 2 groups

t-test



Hotelling's  $T^2$

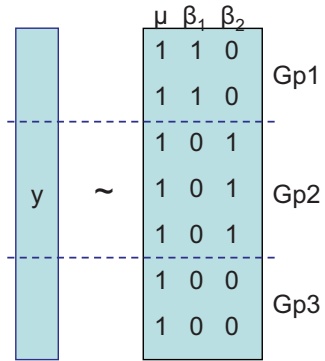


**Multivariate generalization:** find lin. comb. of  $y$ 's  $\rightarrow$  max. univariate  $t^2$ . (Wts are discriminant coefficients.)

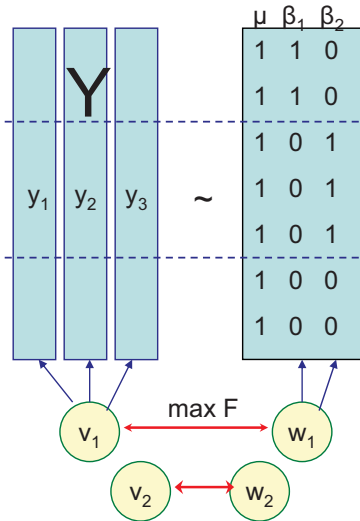
20

# Discrete predictors: 1 factor

1-way ANOVA

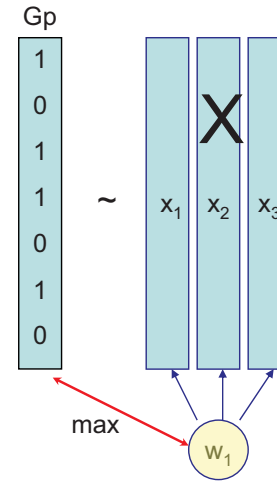


1-way MANOVA



**Multivariate generalization:** find lin. comb. of y's  $\rightarrow$  max. univariate F

# Discrete responses



- **Discriminant analysis:** find lin. comb. of x's that maximally separates groups  $\rightarrow$  max F
- **Logistic regression:** find lin. comb. of x's that maximally predicts  $p \equiv \text{Prob}(y=1)$

Logistic regression as a **generalized** linear model:

$$\log \text{ odds} = \log \left( \frac{p}{1-p} \right) = \mathbf{X}\beta$$

Full generalized linear model for non-normal data:

$$g(y) = \mathbf{X}\beta$$

# Discrete responses & predictors

Job Satisfac

L	M	H
1	0	0
0	1	0
1	0	0
0	1	0
0	0	1
0	1	0
0	1	0
0	0	1
0	0	1

Education

L	M	H
1	0	0
1	0	0
1	0	0
0	1	0
0	1	0
0	1	0
0	0	1
0	0	1
0	0	1

Education (x)

	Lo	M	Hi
L	23	10	5
M	12	37	9
H	4	9	43

Simplest example:  $\chi^2$  for 2-way table

Multi-way frequency tables: **loglinear models** account for associations among discrete factors

$$\log(f) = \mathbf{X}\beta$$

# Techniques, by variable type

Response variables:  $y_1, \dots, y_q$

Predictor variables: $x_1, \dots, x_p$	Quantitative	Quantitative		Discrete	
		q=1	q>1	q=1	q>1
Discrete	p=1	Simple regression	MMRA	Simple logistic regression	
	p>1	MRA	MMRA Canonical corr. Partial corr.	Mult. logistic regression Discriminant analysis	Multivariate logistic regression
Discrete	p=1	t-test 1-way ANOVA	Hotelling T <sup>2</sup> 1-way MANOVA	Simple $\chi^2$	Loglinear models
	p>1	Factorial ANOVA	Factorial MANOVA	Logit models Loglinear models	

## Graphical methods + Geometry=Insight

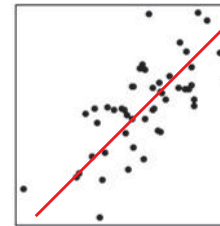
- **Graphical methods:** major theme of this course
  - No data analysis is well-begun or well-completed without extensive, well-chosen data displays
  - **Data analysis = Summarization + Exposure**  
(statistical model) (graphs)
  - **Visual statistics:** Let your data tell you what they seem to say – graphs speak more clearly than a  $p$ -value.
  - **Visual diagnostics:** graphical methods for diagnosing violations of model assumptions & suggesting corrective actions.

25

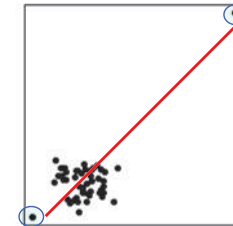
## Visual statistics: Why plot your data?

Three data sets with exactly the same bivariate summary statistics:

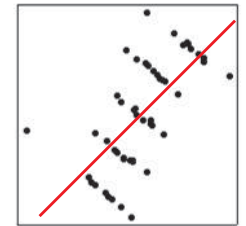
- Same correlations, linear regression lines, etc
- Indistinguishable from standard printed output



Standard data



$r=0$  with 2 outliers



Lurking variable?

26

## Graphical methods + Geometry=Insight

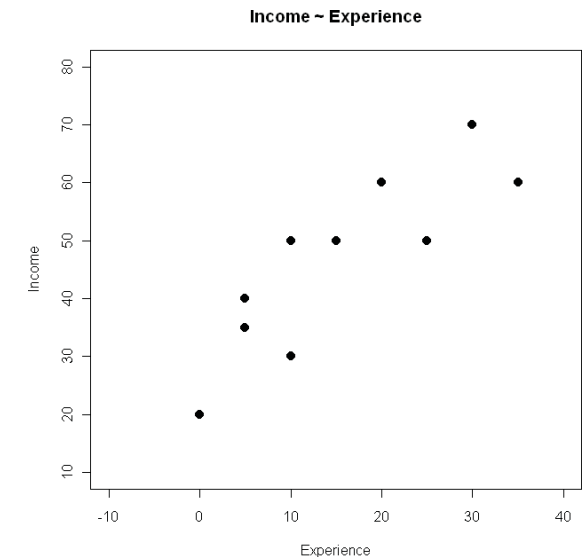
- **Geometry:** visual understanding of statistical concepts
  - Regression: fitting lines, planes, hyperplanes
  - Fitting by least squares: projection of  $\mathbf{y}$  on  $\mathbf{X}$
  - df: # of dimensions of a vector space
  - SS: lengths of vectors
  - Ellipses: visual summaries of data (data ellipses) and models (confidence ellipses)
  - Helps to use 2D (& 3D) to understand high-D data

27

## Geometry: Data ellipse

Looking at scatterplots:

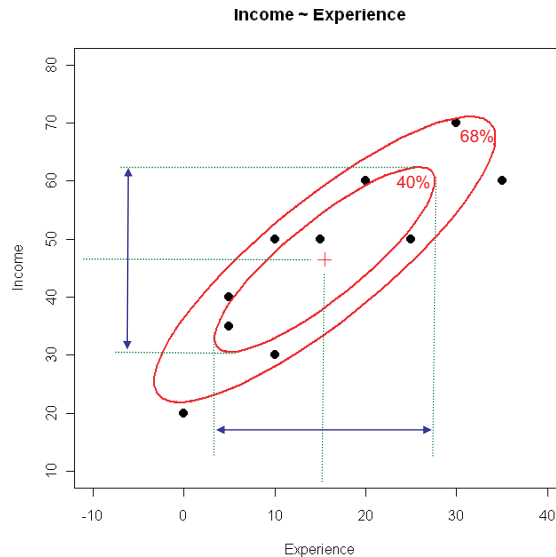
- What is SD of  $x$ ? of  $y$ ?
- What is correlation?
- What is regression line?
- Is relationship linear?
- Are there unusual pts?



## Geometry: Data ellipse

### Data ellipse:

- Encloses  $(1-\alpha)\%$  in bivariate normal dist
- 40% = univariate std interval = mean  $\pm$  1 SD
- 68% = bivariate std interval



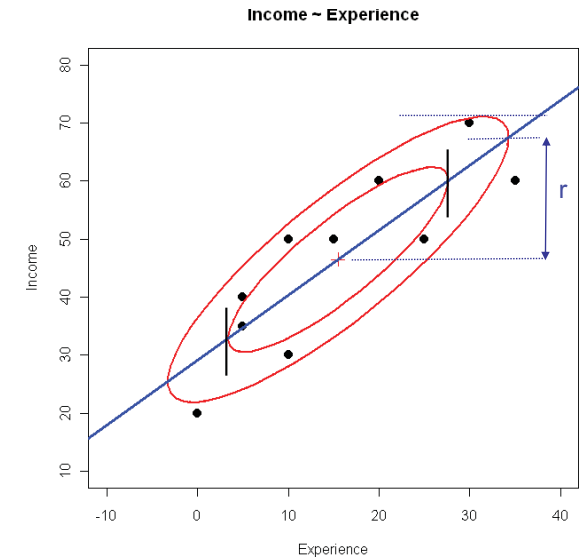
## Geometry: Data ellipse

### Regression & correlation:

- Regression of y on x goes thru pts of vertical tangency
- correlation is the ratio of height of regression line to height of data ellipse
- visual estimates:  

$$\text{Inc} \approx 29 + 1.1 \text{ Exp}$$

$$r \approx 0.85$$



## Summary

- Multivariate analysis unifies all traditional linear models within the GLM framework
- Concepts, statistics, and tests apply equally for regression & ANOVA
- All methods involve linear combinations, optimizing some criterion
- Easy generalizations:
  - Multivariate models:  $\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon} \rightarrow \mathbf{Y} = \mathbf{X} \mathbf{B} + \mathbf{E}$
  - Non-normal data: models for  $g(y)$ 
    - Logistic/logit models:  $\log [p/1-p] = \mathbf{X} \boldsymbol{\beta}$
    - Loglinear models:  $\log(f) = \mathbf{X} \boldsymbol{\beta}$
- Graphical methods + Geometry = Insight!