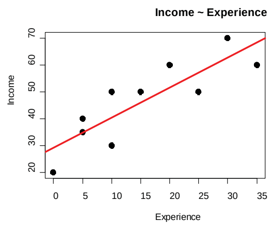
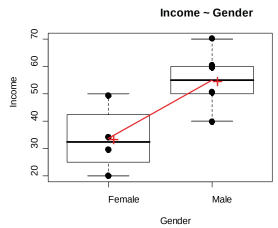
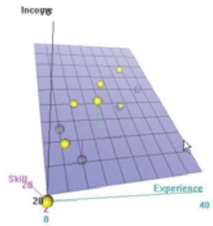
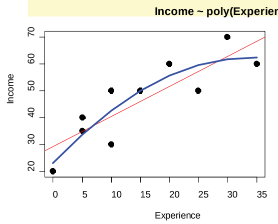




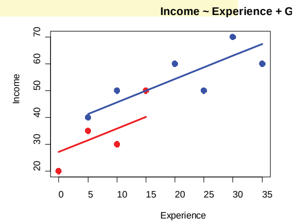


Multivariate Data Analysis: Overview



Michael Friendly

Psychology 6140



Overview of Overview

- Today, I'm going to try to paint an overview of the content of the course with a very broad brush.
- The key ideas are:
  - Linear models (regression, ANOVA) extend directly to multivariate response data
  - Nearly all models involve linear combinations (weighted sums)
  - Models and data can be more easily understood with graphics
  - Statistical ideas have a visual representation in geometry
- Multivariate techniques can be classified by the attributes of
  - data (quantitative vs. categorical)
  - Numbers of predictors and response variables

Why study multivariate data analysis?

- Multivariate data more common in research
- GLM approach: ANOVA, regression, etc. within a common framework: linear models

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i$$

- In matrix form ($\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$), GLM extends to MANOVA, MMReg, etc.
- Idea of linear combinations extends readily to other methods: PCA, discriminant analysis, etc.
- Graphical methods, geometry → Insight

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.

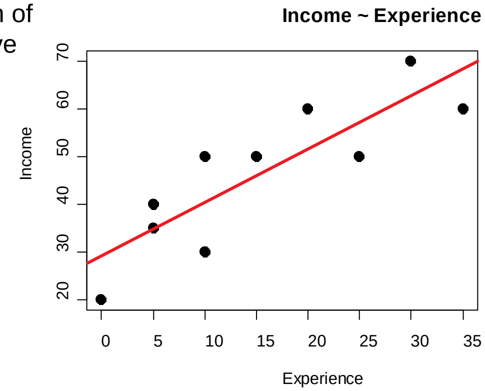
How do these vary with predictors?

1. Linear models: Regression

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

Model: $E(y \mid x) = \beta_0 + \beta_1 x$

e.g, Income = 29 + 1.12 Experience



Parameters:

β_0 = 29 = Income at 0 years

β_1 = 1.12 = Increase / year = $\frac{\Delta y}{\Delta x}$

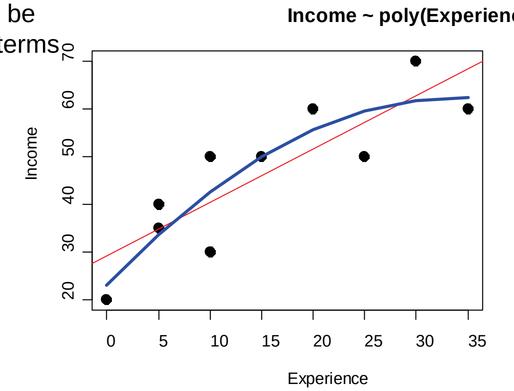
The regression line on the graph, and the fitted equation are just summaries. It is important to think about what they mean for a given problem!

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: $E(y \mid x) = \beta_0 + \beta_1 x + \beta_2 x^2$

e.g, Income = 23 + 2.3 Exp -0.33 Exp^2



What does β_1 = 2.3 mean?

What does $\beta_2 < 0$ mean?

Parameters:

β_0 = 23 = Income at 0 years

β_1 = 2.3 = Slope at 0 years

β_2 = -0.33 = Decrease in slope/year

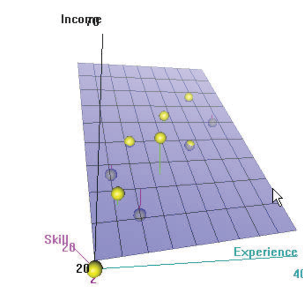
The graph of predicted values gives a visual interpretation

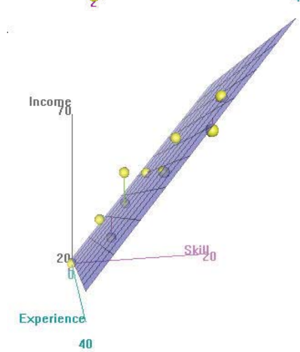
Linear models: Multiple regression

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

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

e.g, Income = 14.8 + 0.11 Exper + 3.4 Skill





Parameters:

β_0 = 14.8 = Income at 0 years, 0 skill

β_1 = 0.11 = ΔIncome /ΔExperience | Skill

β_2 = 3.4 = ΔIncome /ΔSkill | Experience

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

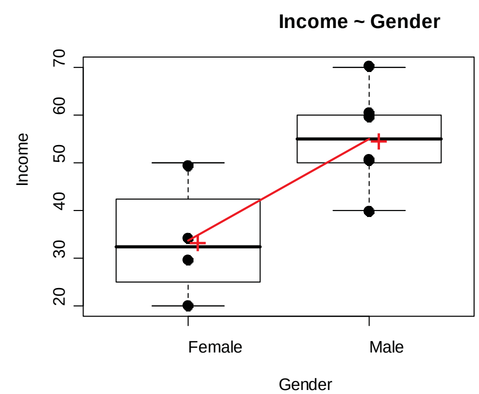
Linear models: ANOVA

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

Model: E(Y) = μ + β (G='Male')

e.g., Income = 33.75 + 21.25 (G='Male')

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



Parameters:

μ = 33.75 = Female mean Income

β = 21.25 = Increment for Male

How would you describe this in words?

Linear models: Regression + Anova

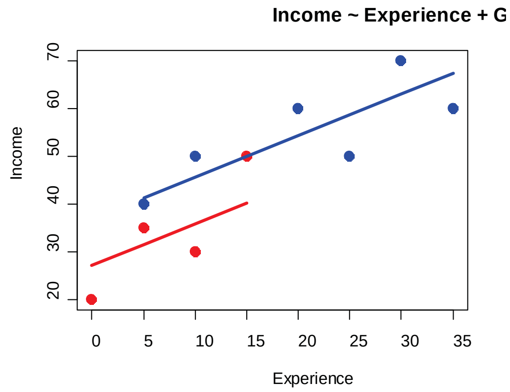
ANCOVA: Is there a difference in a factor, controlling for a quantitative predictor?

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.,

Inc = 27.27 + 0.86 Exp + 9.73 (G='Male')



The coefficient, β_2 for G='Male' allows the intercepts (or means) to differ. Slopes are forced to be equal.

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')

$+ \beta_3$ X1 * (G='Male')

e.g.,

Inc = 21.0 + 1.70 Exp + 19.25 (G='Male')

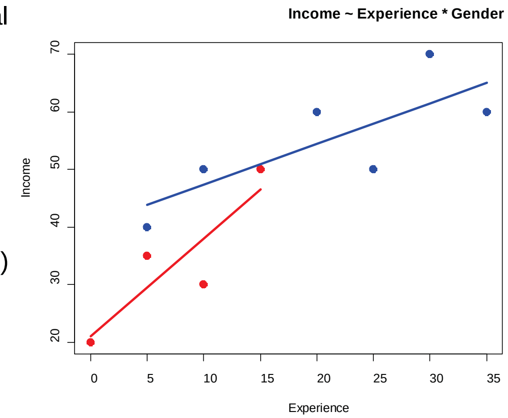
- 1.0 Exp * (G='Male')

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



A more complete description, but maybe overly complex!

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

Regression and ANOVA are basically the same model, but use different terminology and emphasize different stats

General Linear Model (GLM)

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

$$\underbrace{\overbrace{y_i}^{\text{Outcome}}}_{\text{data}} = \underbrace{\overbrace{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}}^{\text{linear combination of predictors}}}_{\text{fitted (explained part)}} + \underbrace{\overbrace{\varepsilon_i}^{\text{residual}}}_{\text{residual (unexplained)}}, \quad \varepsilon_i \sim N(0, \sigma^2)$$

where,

	Regression	ANOVA
X	Quantitative predictor (experience, skill)	Indicator (0/1) variables for group membership
β	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:

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

or,

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times(p+1)}\boldsymbol{\beta}_{(p+1)\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$$

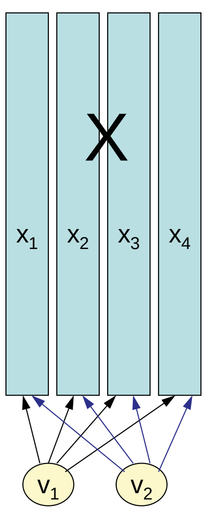
For all cases:

- 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.

2. 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

Linear combinations: 1 set of variables

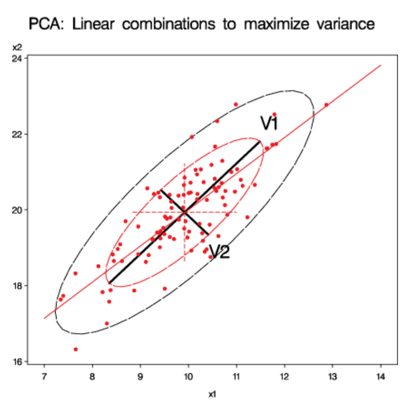


PCA: find weights to maximize variance of v_1, v_2, …

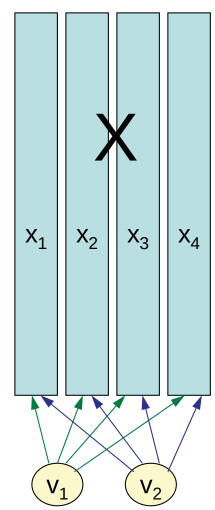
$v_1 = a_1\, x_1 + a_2\, x_2 + a_3\, x_3 + a_4\, x_4$

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

subject to: all v_i, v_j uncorrelated

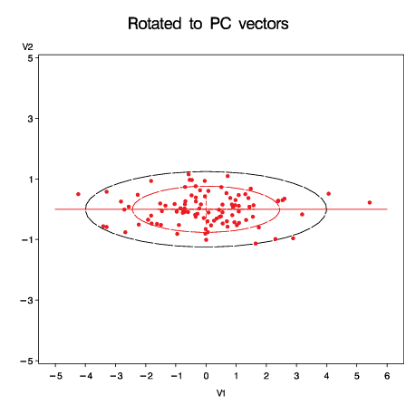


Linear combinations: 1 set of variables

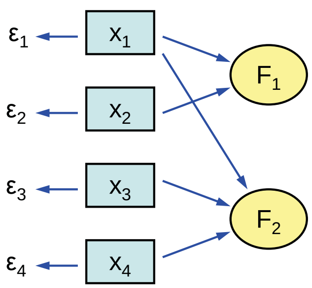


With *p* variables, *p* components account for 100% of variance, and correspond to a rotation of the variable space to uncorrelated components.

Goal in PCA is to account for most variance with k<<p components.



Factor analysis: Latent variables



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

$x_1 = \lambda_{11} F_1 + \lambda_{12} F_2 + \varepsilon_1$

$x_2 = \lambda_{21} F_1 \qquad + \varepsilon_2$

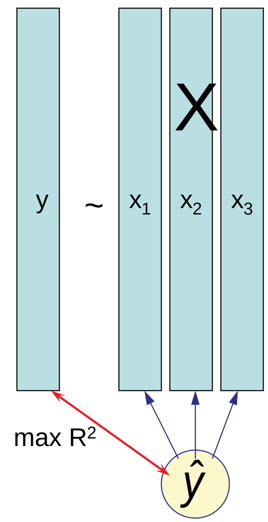
$x3 = \qquad \lambda_{32} F_2 + \varepsilon_3$

$x4 = \qquad \lambda_{42} F_2 + \varepsilon_4$

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

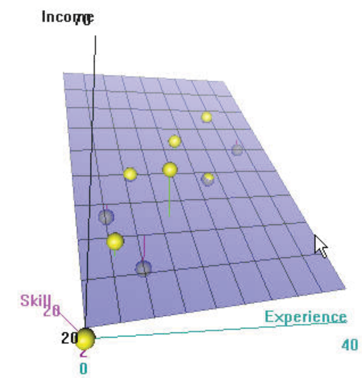
Linear combinations: 2 sets of variables



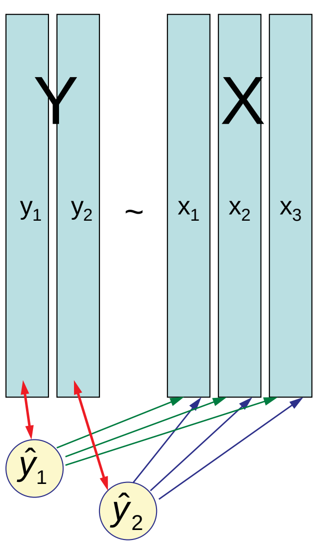
Univariate response:

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

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$$



2 sets, multivariate response: MMRA



Multivariate response: MMRA

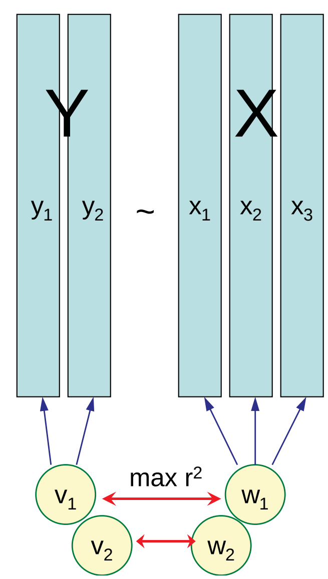
Multivariate MRA: find weights to maximize correlation between *each* y and predicted y,

$$\hat{y}_1 = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$$

$$\hat{y}_2 = c_0 + c_1 x_1 + c_2 x_2 + c_3 x_3$$

- 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."

2 sets, multivariate response: CanCorr



Canonical correlation:

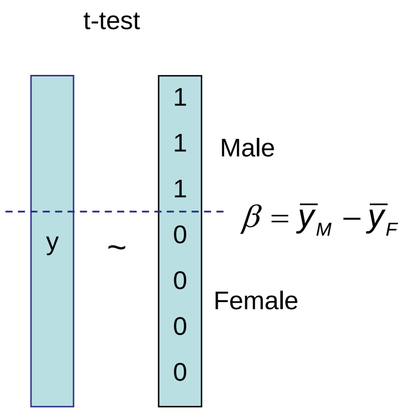
Find linear combinations of the x's that best predicts linear combination of the y's

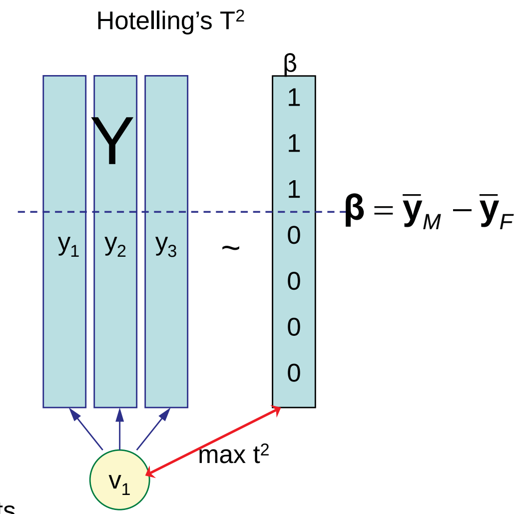
$v_1 = a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 x_4$

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

- Choose weights to maximize r^2 (v1,w1)
- Up to s=min(p,q) additional pairs of canonical variables: (v_2, w_2), … (v_s, w_s)
- All correlations between the Ys and Xs are explained thru the correlation of each v_i with w_i.

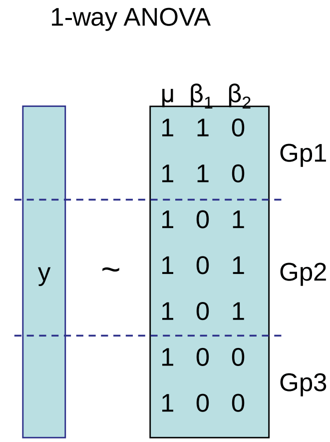
Discrete predictors: 2 groups

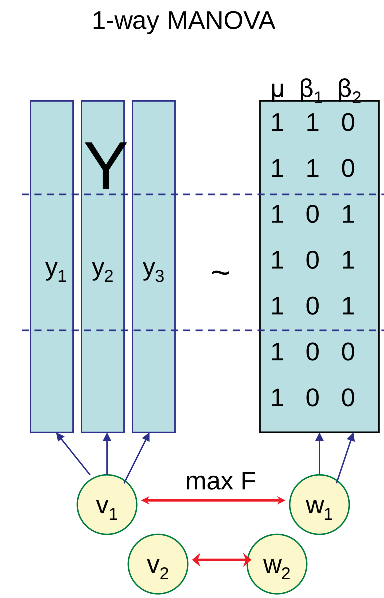




Multivariate generalization: find lin. comb. of y's → max. univariate t^2. (Wts are discriminant coefficients.)

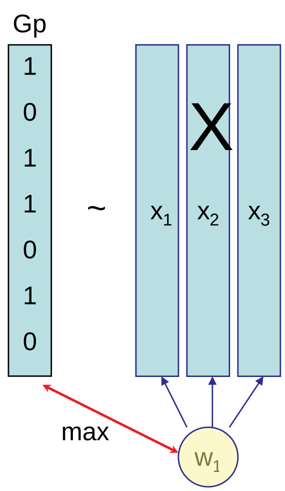
Discrete predictors: 1 factor





Multivariate generalization: find lin. comb. of y's → max. univariate F

Discrete responses



- **Discriminant analysis**: find lin. comb. of x's that maximally separates groups → max F
- **Logistic regression**: find lin. comb. of x's that maximally predicts p Ξ Prob(y=1)

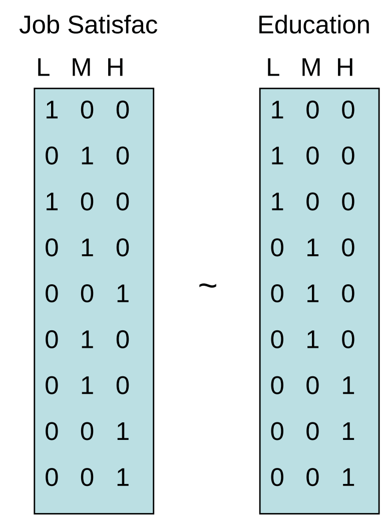
Logistic regression as a **generalized** linear model:

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

Full generalized linear model for non-normal data:

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

Discrete responses & predictors



Education (x) / Satisfaction (y)

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

Simplest example: χ2 for 2-way table

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

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

Techniques, by variable type

Response variables: $y_1, \ldots y_q$

Predictor variables: $x_1, \ldots x_p$

		Quantitative		Discrete	
		q=1	q>1	q=1	q>1
Quantitative	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^2 1-way MANOVA	Simple χ^2	Loglinear models
	p>1	Factorial ANOVA	Factorial MANOVA	Logit models Loglinear models	

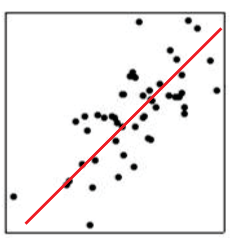
3. Graphical methods + Geometry=Insight

- **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.

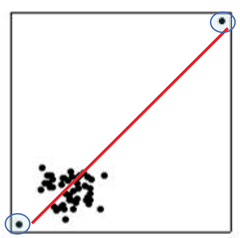
Visual statistics: Why plot your data?

Three data sets with exactly the same bivariate summary statistics:

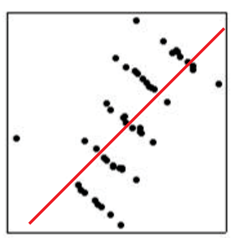
- Same correlations, linear regression lines, etc
- Indistinguishable from standard printed output



Standard data



r=0 but + 2 outliers



Lurking variable?

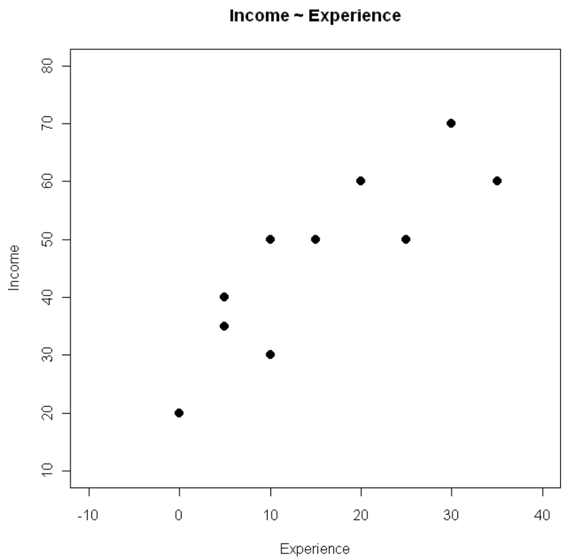
Graphical methods + Geometry=Insight

- **Geometry**: visual understanding of statistical concepts
  - Regression: fitting lines, planes, hyperplanes
  - Fitting by least squares: projection of ***y*** on ***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

Geometry: Data ellipse

Looking at scatterplots:

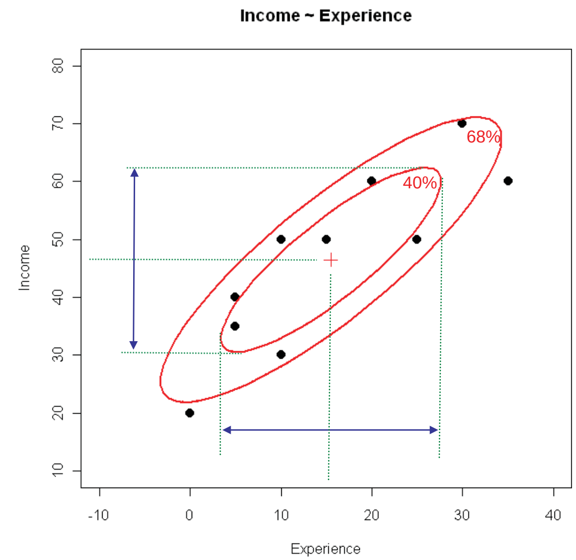
- What is SD of x? of y?
- What is correlation?
- What is regression line?
- Is relationship linear?
- Are there unusual pts?



Geometry: Data ellipse

Data ellipse:

- Encloses (1-α)% in bivariate normal dist
- 40% = univariate std interval = mean ± 1 SD
- 68% = bivariate std interval



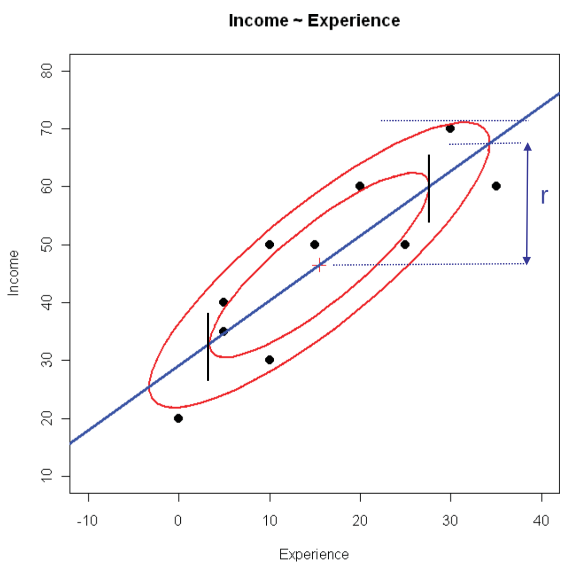
Geometry: Data ellipse

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:

Inc ≈ 29 + 1.1 Exp

r ≈ 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: $\mathbf{y = X \beta + \varepsilon \rightarrow Y = X B + E}$
  - Non-normal data: models for g(y)
    - Logistic/logit models: log [p/1-p] = $\mathbf{X \beta}$
    - Loglinear models: log(f) = $\mathbf{X \beta}$
- Graphical methods + Geometry = Insight!