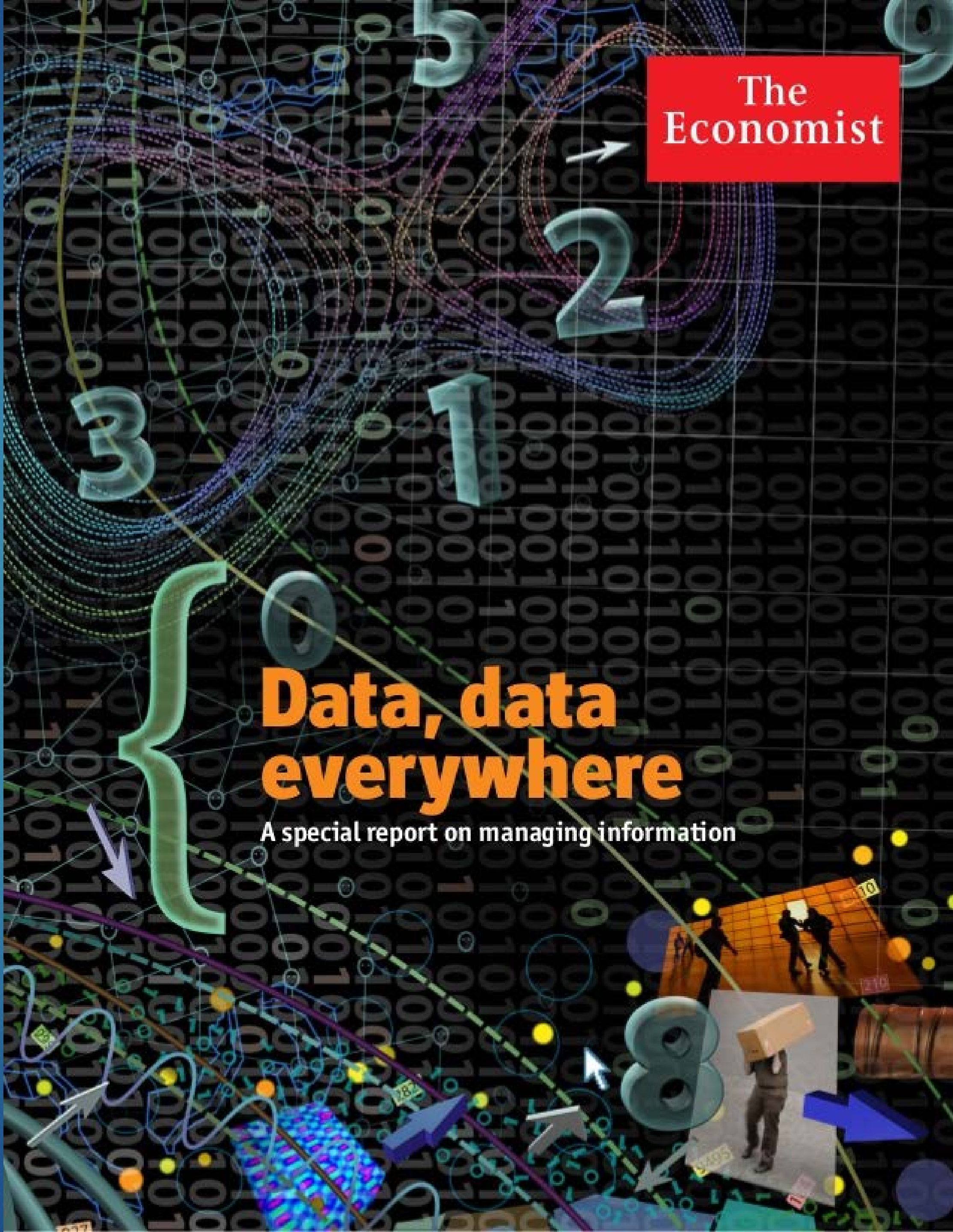
ECE3340

Numerical Fitting, Regression, Interpolation and Approximation

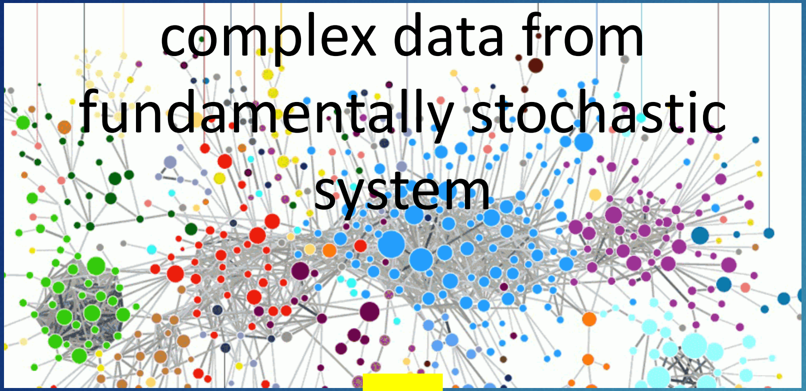
PROF. HAN Q. LE

Note: PPT file is the main outline of the chapter topic – associated Mathematica file(s) contain details and assignments

Overview

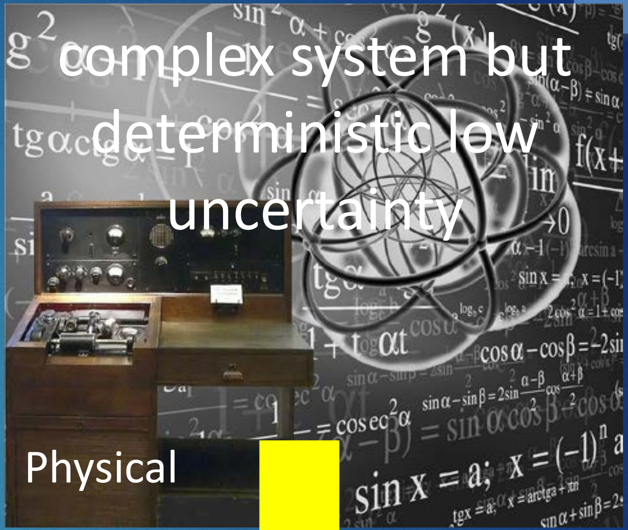

The Economist
Data, data everywhere
A special report on managing information

Types of problems



complex data from fundamentally stochastic system

Medical, social



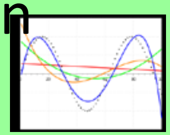
complex system but deterministic low uncertainty

Physical

simplified phenomenological model

$f(x_1, x_2 \ldots , t)$

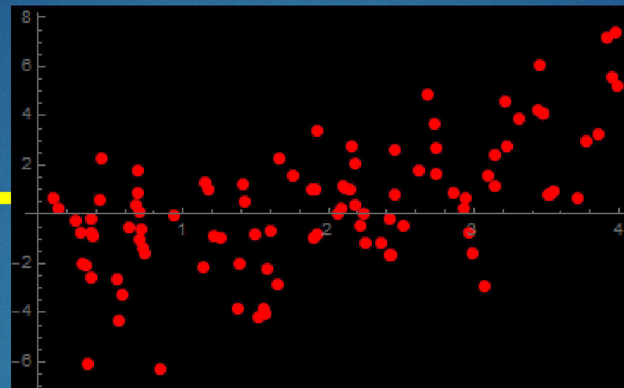
simplified, fitting, approximation



The underlying motivation is to simplify - to make things easy to understand or to implement

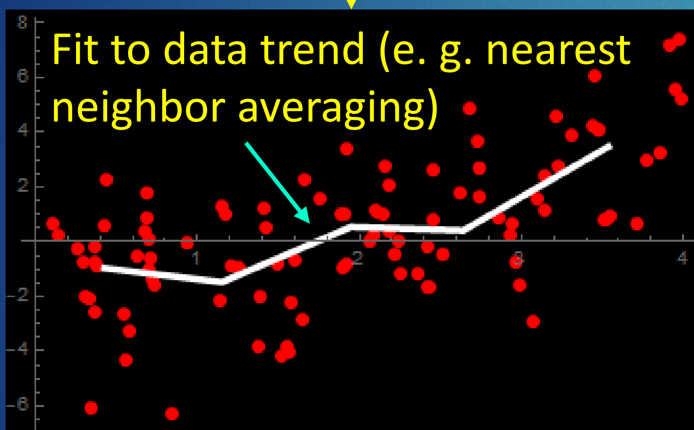
A crucial fundamental difference

Intrinsically stochastic – large fluctuation is intrinsic – not nec. errors



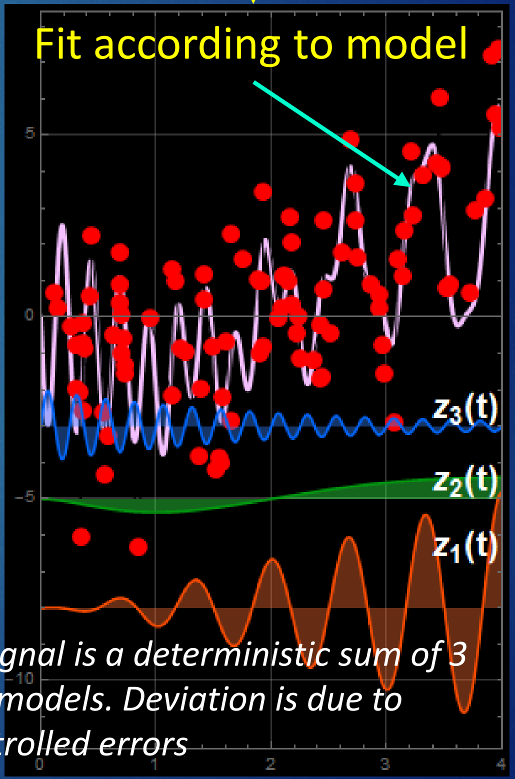
Deterministic basis (e. g. per physical law) with known model + random errors

Data Analytics



Example: a medicine efficacy: fluctuation is not due measurement errors, but genuine individual genetic/ lifestyle variation.

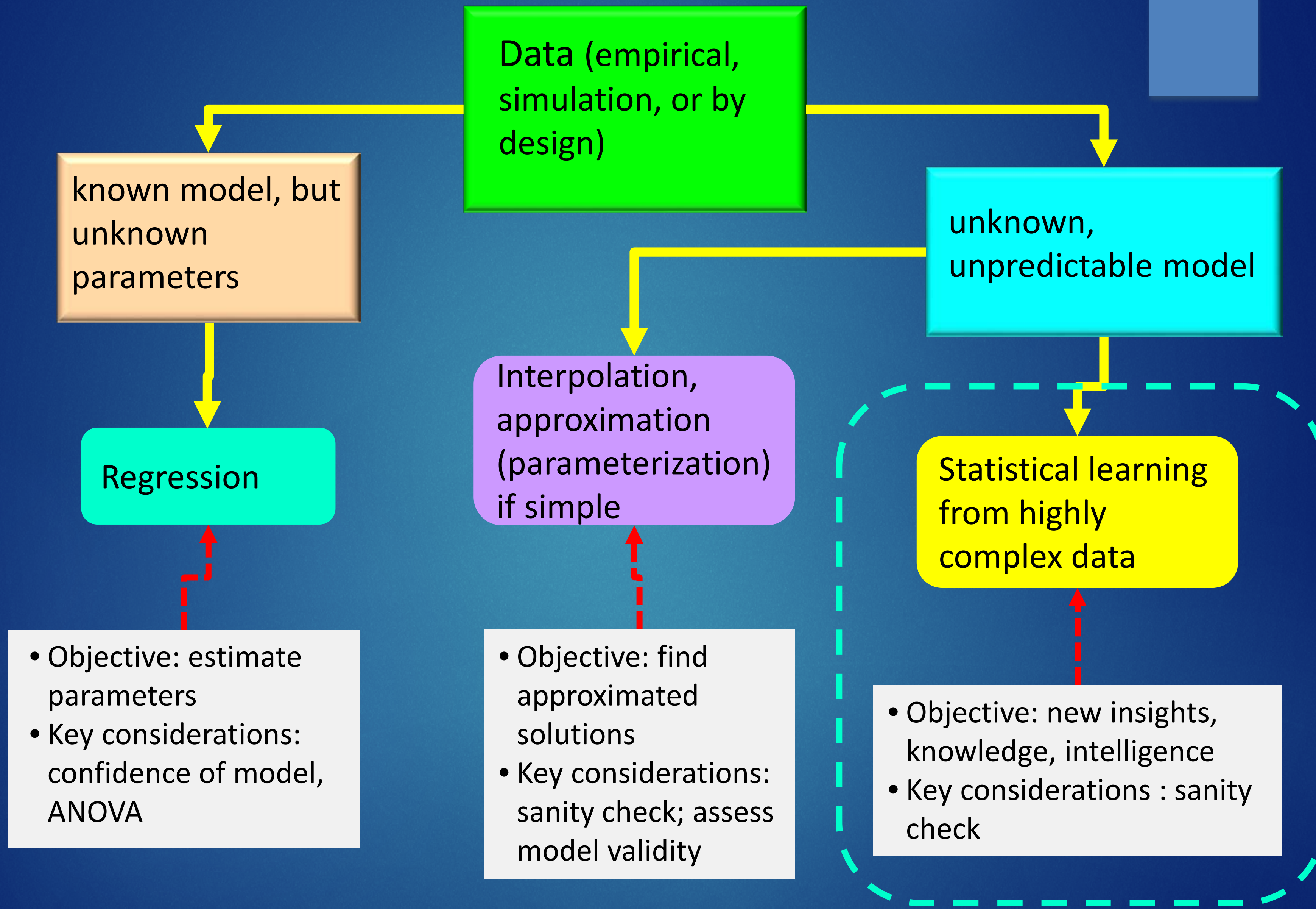
Model Analytics



Example: a signal is a deterministic sum of 3 known basis models. Deviation is due to small, uncontrolled errors

Part A – Data Analytics

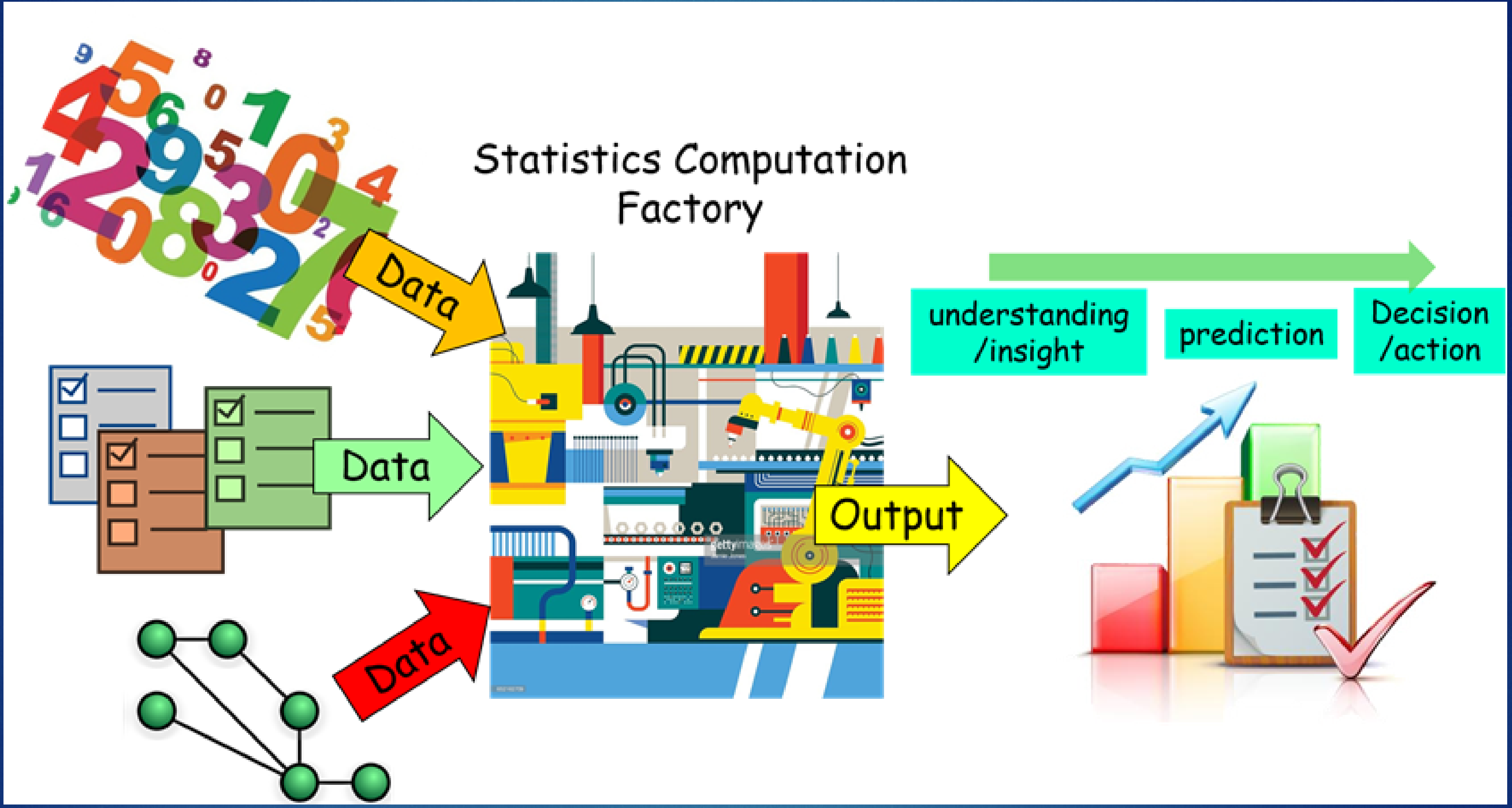
DATA PRESENTATION (VISUALIZATION), STATISTICS, REGRESSION, CLASSIFICATION.


Data (empirical, simulation, or by design)
known model, but unknown parameters
unknown, unpredictable model
Regression
Interpolation, approximation (parameterization) if simple
Statistical learning from highly complex data
• Objective: estimate parameters
• Key considerations: confidence of model, ANOVA
• Objective: find approximated solutions
• Key considerations: sanity check; assess model validity
• Objective: new insights, knowledge, intelligence
• Key considerations : sanity check


AI


BIG DATA

Data Analytics – Statistical Learning



This has evolved from data analysis to data analytics, pattern classification

Outline

- Linear regression
  - single variable
  - multiple variables
- Linearization of non-linear model
  - linear-exponential or log-linear
  - power-relationship: log-log
  - general non-linear
- General model fit
  - Least squares of linear combination of basis functions
- Discrete (quantized) variables and generalized regression: logistic regression
- Introduction to data clusters and classification

Linear Regression

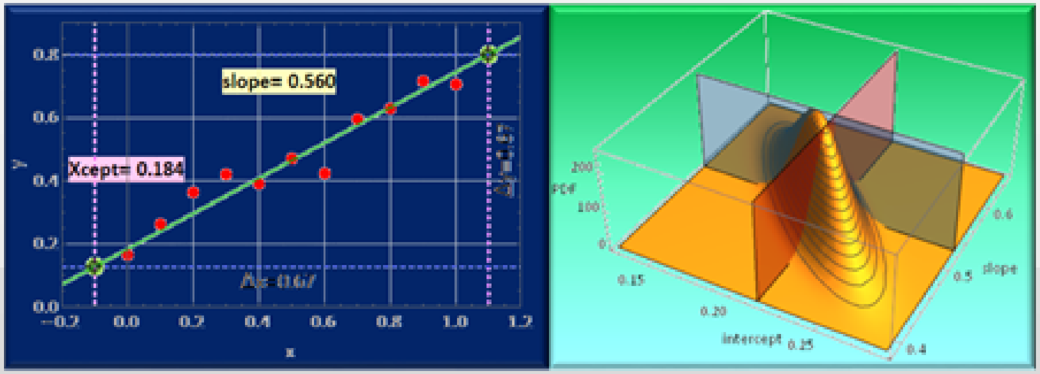
Introduction to regression concept
(see Mathematica lecture file)

Linear regression with single variable

UNIVERSITY of HOUSTON App by Han Q. Le ©

ECE3340–APP–Linear regression

RUN STATUS/CONTROL

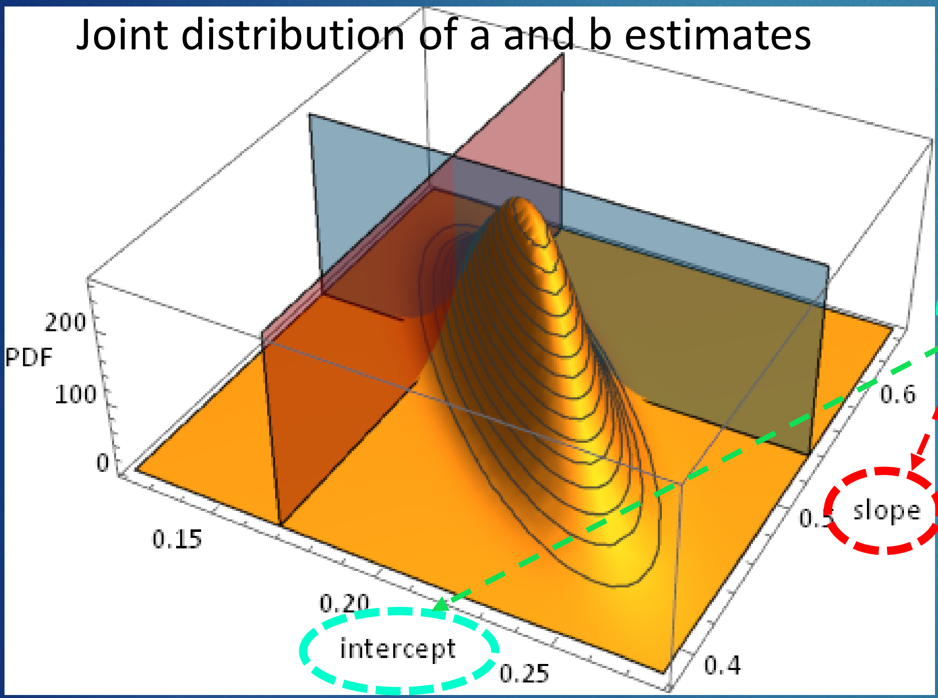


	Deg. freed.	Sum sq	Mean sq	F–Statistics	P–Value
x	1	0.296605	0.296605	133.17	1.07351×10^{-6}
Error	9	0.0200454	0.00222727		
Total	10	0.31665			

Key concepts

- Model parameters:
  - coefficients, correlation R2
  - standard error, covariance matrix, correlation matrix
  - confidence ellipsoid
- Linear regression statistics
  - residuals
  - parameter t-statistics, P-value
  - Analysis of variance (ANOVA): dof, sum of squares, mean squares, F-statistics

Covariance matrix of parameters – Confidence ellipsoid



$$\hat{a} = \frac{\frac{\sum_{i=1}^{n} x_i\, y_i}{n} - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n^2}}{\frac{\sum_{i=1}^{n} x_i{}^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2} = \frac{\overline{x\,y} - \overline{x}\;\overline{y}}{\overline{x^2} - \overline{x}^2}$$

$$\hat{b} = \overline{y} - \hat{a}\,\overline{x}$$

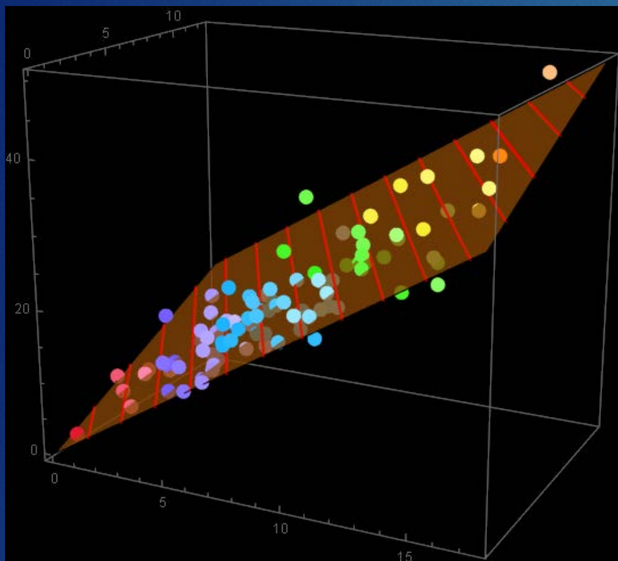
Estimates for a and b are not independent. They are related by mean x and mean y as shown, hence, the distribution of their values are not independent.

in class demo: if we know one coefficient by any other mean, this changes the estimate for the other coefficient (move the planes).

Key concepts

- Model parameters:
  - coefficients, correlation R2
  - standard error, covariance matrix, correlation matrix
  - confidence ellipsoid
- Linear regression statistics
  - residuals
  - parameter t-statistics, P-value
  - Analysis of variance (ANOVA): dof, sum of squares, mean squares, F-statistics

From the example discussed



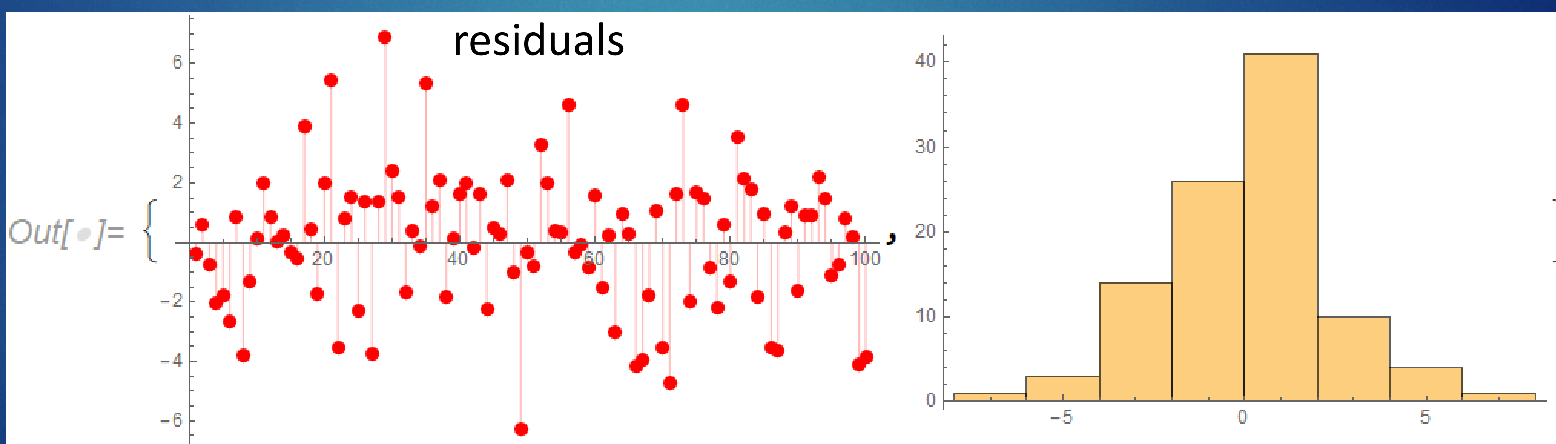
Null hypothesis

Out[ ]=

	Estimate	Standard Error	t-Statistic	P-Value
u1	2.03334	0.0463809	43.8399	3.34417×10^{-66}
u2	1.1749	0.0853734	13.7619	1.25441×10^{-24}

Out[ ]=

	DF	SS	MS	F-Statistic	P-Value
u1	1	48 363.1	48 363.1	8665.47	1.93448×10^{-97}
u2	1	1057.02	1057.02	189.391	1.25441×10^{-24}
Error	98	546.951	5.58113		
Total	100	49 967.1			



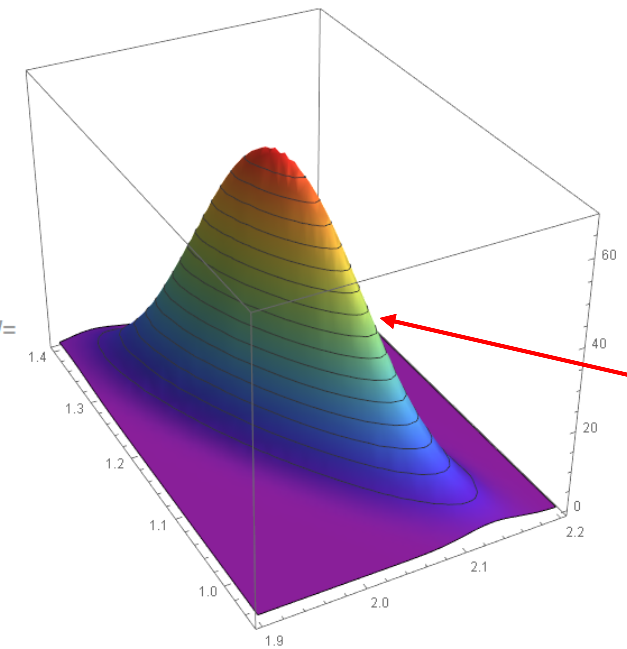
```
In[•]:= covMat // MatrixForm
```

Out[•]//MatrixForm=

$$\begin{pmatrix} 0.00215119 & -0.00319625 \\ -0.00319625 & 0.00728862 \end{pmatrix}$$

```
In[•]:= Plot3D[PDF[MultinormalDistribution[ahat, covMat], {u, v}],
        {u, 1.9, 2.2}, {v, 0.95, 1.4}, PlotRange → All
        , BoxRatios → {3, 4.5, 3}, ColorFunction → "Rainbow"
        , MeshFunctions → {#3 &}]
```

Out[•]=




```
In[•]:= eigen
```

Out[•]=

Eigenvalue	Index	u1	u2
1.10248	1.	0.448759	0.448759
0.897518	1.10832	0.551241	0.551241

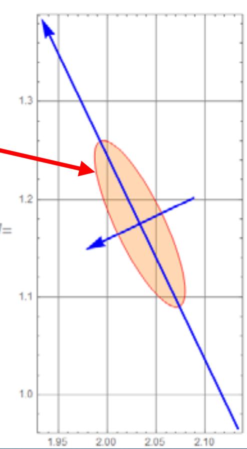
```
In[•]:= ahatconf
```

Out[•]=

	Estimate	Standard Error	Confidence Interval
u1	2.03334	0.0463809	{1.9413, 2.12538}
u2	1.1749	0.0853734	{1.00548, 1.34432}

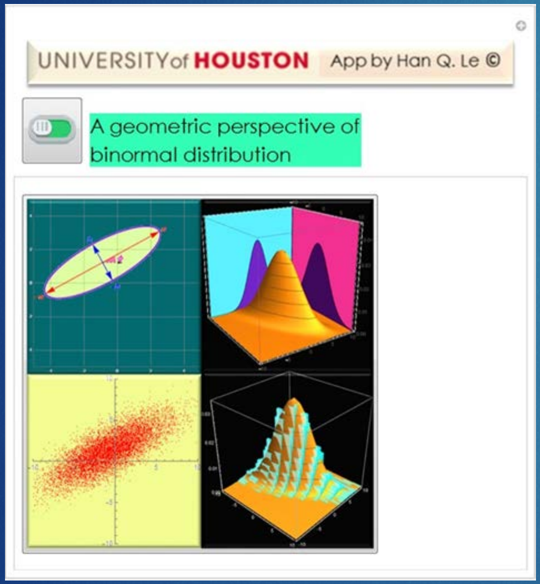
```
Out[•]= FittedModels`ParameterEllipsoid[
        {2.03334, 1.1749}, {0.233444, 0.0618608},
        {{-0.432183, 0.901786}, {-0.901786, -0.432183}}]

In[•]:= Graphics[{EdgeForm[{Thickness[0.005], Red}],
         Hue[0.08, 1, 1, 0.3], Ellipsoid[ahat, covMat]
         , Blue, Opacity[1], Thickness[0.01], Arrowheads[0.1]
         , Arrow[#] & /@
          {({ahat, ahat} + ahatellipse[[2, 1]] *
              {-ahatellipse[[3, 1]], ahatellipse[[3, 1]]})
           , ({ahat, ahat} + ahatellipse[[2, 2]] *
              {-ahatellipse[[3, 2]], ahatellipse[[3, 2]]})}}
        , Frame → True, GridLines → Automatic]
```

Out[•]=

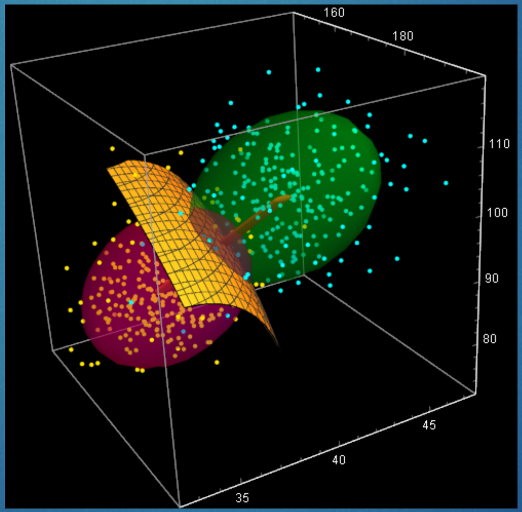


A very useful exercise to understand variable correlation

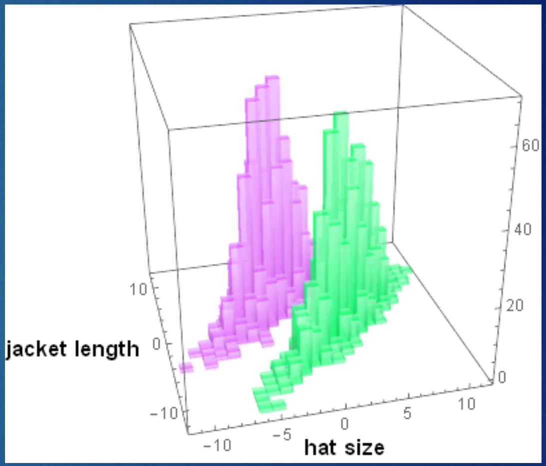
Review of binormal distribution



Example: human body measures- gender correlation



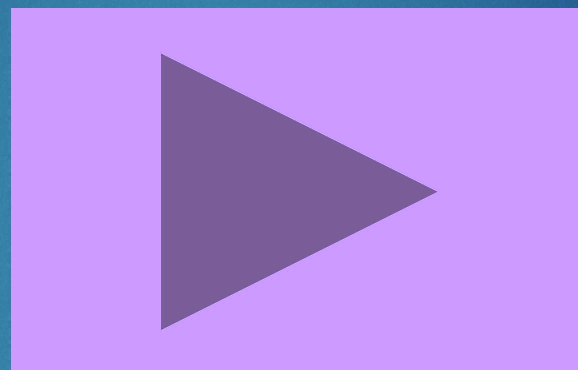
Example in HW



Outline

- Linear regression
  - single variable
  - multiple variables
- Linearization of non-linear model
  - linear-exponential or log-linear
  - power-relationship: log-log
  - general non-linear
- General model fit
  - Least squares of linear combination of basis functions
- Discrete (quantized) variables and generalized regression: logistic regression
- Introduction to data clusters and classification

Mathematica file on generalized least-square regression



Introduction to data survey and visualization methods
(see Mathematica lecture file)