Modelado y Simulación

Julián Jiménez Cárdenas

Facultad de Matemáticas
FUKL

August 11, 2024

Introduction to Modeling I

Mathematical modeling is a powerful tool that:

- Translates real-world problems into mathematical language
- Provides insights into complex systems
- Enables predictions and decision-making
- Facilitates optimization and efficiency

Key applications include:

- Physics and engineering
- Economics and finance
- Biology and medicine
- Climate science and ecology

Elaborating in some applications, we have

Introduction to Modeling II

- **Celestial Mechanics**
  - Understanding motion of stars, planets, comets
  - Applications in agriculture, navigation, calendars
  - Modern astrophysics: universe formation and development
- **Energy Supply**
  - Modeling energy consumption trends
  - Planning for future energy needs
- **Everyday Scenarios**
  - E.g., safe driving distances at various speeds
- **Climate Change**
  - Modeling global warming impacts
  - Predicting changes in weather patterns
  - Assessing human impact on climate

Introduction to Modeling III

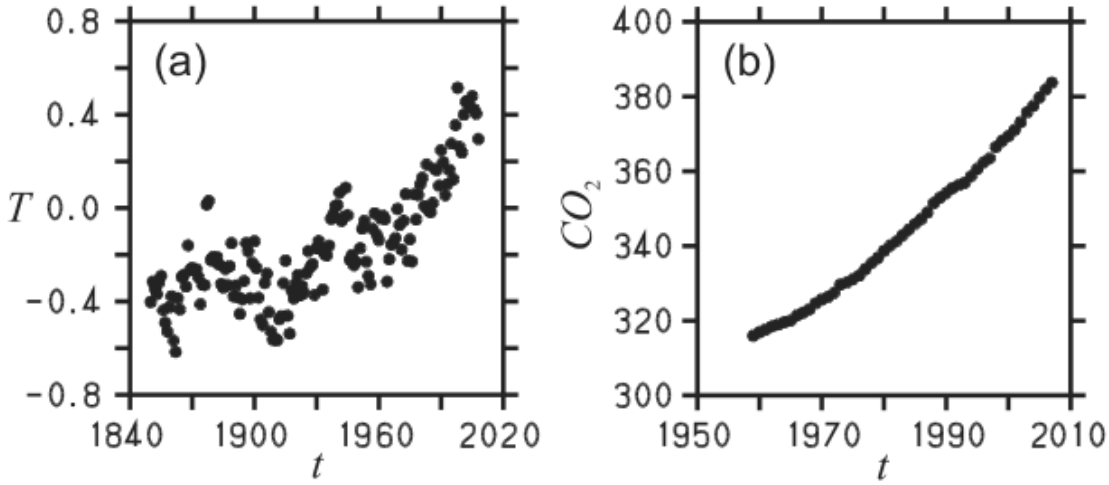

Figure: An illustration of global warming. (a) The HadCRUT3 global temperature anomaly data (consisting of annual differences from 1961-90 normals) are given in °C (Rayner et al, 2003, Brohan et al, 2006); (b) the Mauna Loa data (Tans 2008) of atmospheric CO_2 concentrations are given in ppmv.

Linear Models I

First we introduce the simplest modeling approach: **the linear functions**.

- Suppose we aim to set a relation between two variables, **x** and **y**, and we have paired data from these variables of the form

$$\mathcal{D} = \{(x_i, y_i) \ : \ i = 1, 2, \ldots, n\}.$$

- Thus, we can choose any pair of these points $(x_i, y_i), (x_j, y_j) \in \mathcal{D}$ with $i \neq j$ to find the line that passes through these, which algebraically can be written as

$$y(x) = y_i \frac{x - x_j}{x_i - x_j} + y_j \frac{x - x_i}{x_j - x_i}.$$

- Note that $y(x_i) = y_i$ and $y(x_j) = y_j$, as expected.

Linear Models II

Example (U.S. energy comsumption)

The following table relates the U.S. energy consumption C (in 10^15 Btu) in time t (U.S. Dept, of Energy 2008).

t	C	t	C
1950	34.616	1980	78.122
1955	40.208	1985	76.491
1960	45.087	1990	84.652
1965	54.017	1995	91.173
1970	67.844	2000	98.975
1975	71.999	2005	100.506

- Fit a line passing through the data.
- Graph the error $e = (y - y^{(mod)})/y^{(mod)}$ of the line with respect to the real data. $y^{(mod)}$ is the value predicted by the model.

Exponential Model

Example (Orbital period and distance)

Consider the following data.

Table: Orbital periods and mean distances of planets from the Sun (World Almanac 2010). Here r is the mean distance from the sun in 10^9 km. T_p is the period in earth years $a = 365.25$ days.

Planet	Mercury	Venus	Earth	Mars	Jupiter	Saturn	Uranus	Neptune
r	0.0579	0.1082	0.1496	0.2280	0.7785	1.4335	2.8718	4.4948
T_p	0.2408	0.6152	1.0000	1.8808	11.8618	29.4566	84.0107	164.7858

- Fit a straight line for this data. Is it adequate?
- Assume $T_p = Ar^B$ and use logarithms to transform this relation to a linear one. Fit a straight line. Is it adequate?

Polynomial methods I

The linear approach is useful because of its simplicity, but it has a rather limited range of applicability, as most models cannot be developed in terms of linear functions. Thus, we will consider polynomial models.

- **Linear Polynomials**. There is a unique line passing through a pair $(x_1, y_1), (x_2, y_2)$ of points. Algebraically, let $y = a_0 + a_1 x$ be this line. The requirement of this line passing throught these points imply the following relations

$$y_1 = a_0 + a_1 x_1, \quad y_2 = a_0 + a_1 x_2.$$

This system can be written as

$$\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix},$$

Polynomial methods II

and this can be inverted (almost always) in order to find a_0, a_1, *a.k.a.*, the line, which is

$$P_1(x) = y_1 \frac{x - x_2}{x_1 - x_2} + y_2 \frac{x - x_1}{x_2 - x_1}.$$

- **Non-Linear Polynomials**. Now, if we have a set $\mathcal{P} = \{(x, y_0), (x_1, y_1), \ldots, (x_n, y_n)\}$ of $n + 1$ points, there is (almost always) a polynomial of degree n passing through these points, and it has the form

$$P_n(x) = y_0 L_0(x) + y_1 L_1(x) + \cdots + y_n L_n(x),$$

with

$$L_k(x) = \prod_{i=0, i \neq k}^{n} \frac{x - x_i}{x_k - x_i}.$$

Polynomial methods III

One of the main advantages of exact polynomial models are that they are easily differentiated and integrated. Nevertheless, they are often not very useful, because slight deviations of the data tend to change the trend of the polynomial.

The solution to this problem usually implies working with reduced polynomial models, like **linear models**.

Polynomial methods IV

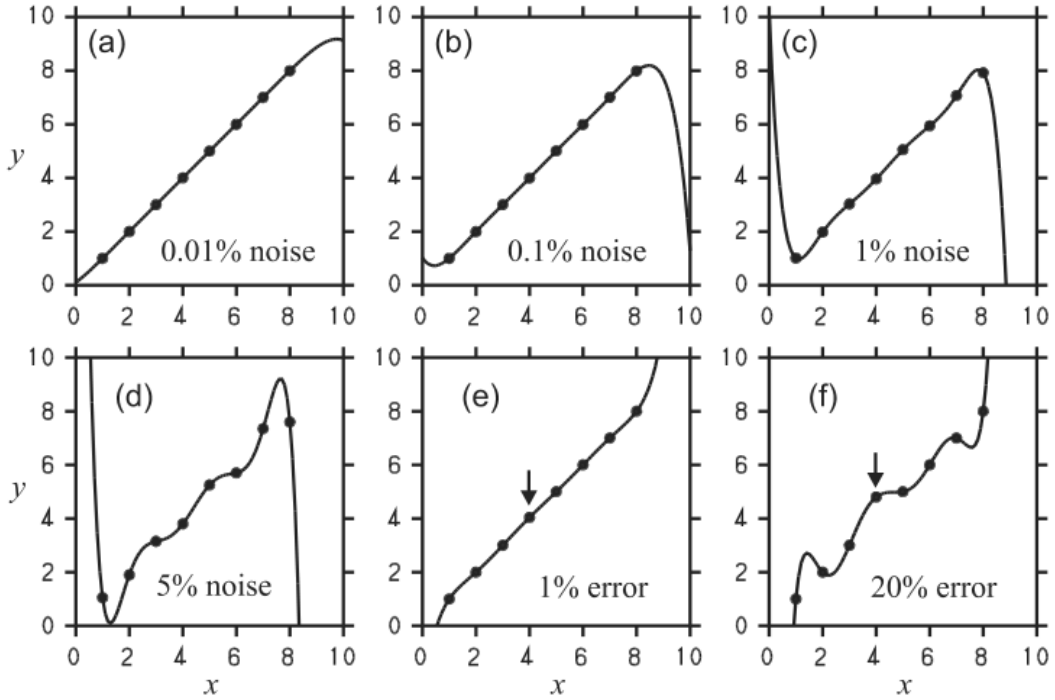

Figure: Exact polynomial models of 7th order for six cases of data points. The data points are given by dots. The polynomial models are given by lines. All the curves pass exactly through the data points. The arrows in (e) and (f) indicate the position of incorrect values.

Polynomial methods V

Consider the development of the world population in time from 1804–2050 according to the Decennial Censuses, U.S. Census Bureau, U.S. Dept. of Commerce (World Almanac 2010). The population P is measured in 10^9 and t refers to the year. The last two population values are projections.

t	1804	1927	1960	1974	1987	1999	2009	2025	2050
P	1.0	2.0	3.0	4.0	5.0	6.0	6.77	7.95	9.32

a) Use the data from 1804 to 2009 to define an exact polynomial of sixth order. Graph this polynomial and the data. Comment on the suitability of this model.

b) Use the data from 1960 to 2009 to define an exact polynomial of fourth order. Graph this polynomial and the data. Comment on the suitability of this model.

Polynomial methods VI

c) Use the data at 1960, 1987, and 2009 to define a polynomial of second order. Graph this polynomial and the data. Comment on the suitability of this model.

d) Use the 1960 and 2009 data to define a polynomial of first order. Graph this polynomial and the data. Comment on the suitability of this model.

The Best Line I

Now that we are again interested in lines, we look for the **best possible line** fitting a set of data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i) \ : \ i = 1, 2, \ldots, n\}$, with $\boldsymbol{x}_i \in \mathbb{R}^m$. Such a line should minimize the error function

$$f = (\boldsymbol{y} - \boldsymbol{y}^{(mod)})^T (\boldsymbol{y} - \boldsymbol{y}^{(mod)}),$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)$ is the vector of the dependent variable, and $\boldsymbol{y}^{(mod)}$ is the vector of predicted values. Now, we take a linear model, so that

$$\boldsymbol{y}^{(mod)} = X\boldsymbol{a},$$

where $X \in M_{n\times(m+1)}$ has in each row the $\boldsymbol{x}_i$'s with a one appended (accounting for the bias term), while $\boldsymbol{a} \in \mathbb{R}^{m+1}$ is the vector of parameters of the model.

The Best Line II

Then, we can show that the best selection of parameters (minimizing f) satisfies

$$a = \left(X^T X\right)^{-1} X^T \boldsymbol{y}.$$

And how do we test if our linear model is accurate? We expect that each component of

$$\boldsymbol{e} = \left(\boldsymbol{y} - \boldsymbol{y}^{mod}\right) / \boldsymbol{y}^{(mod)}$$

is normally distributed around 0 with a "small" variance, *i.e.*,

$$\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2 I).$$

Other situations may indicate that the best model is not linear, and we may need to find other (composite) variables to explain the data tendency.

The Best Line III

Returning to the optimal selection of parameters, the naive complexity order of calculating

$$a = \left(X^T X\right)^{-1} X^T \boldsymbol{y}$$

is $\mathcal{O}\left((m+1)(n+m+1)^2\right)$, which is generally unbearable when either n or m is large. Then, one generally uses numerical methods like **Gradient Descent.** Simply follow the direction of

$$-\nabla_a f = 2X^T \left(\boldsymbol{y} - Xa\right).$$

The algorithm is as follows.

Require: Training data X, target values $\boldsymbol{y}$, learning rate α, number of iterations n_{iters}, tolerance ϵ

Ensure: Optimized parameters a

The Best Line IV

1: Initialize a randomly or with zeros
2: $n \leftarrow$ number of training examples
3: $prev\_cost \leftarrow \infty$
4: **for** $i = 1$ to n_{iters} **do**
5: $\quad h \leftarrow Xa$ ▷ Compute predictions
6: $\quad gradient \leftarrow \frac{1}{n} X^T (h - y)$ ▷ Compute gradient
7: $\quad a \leftarrow a - \alpha \cdot gradient$ ▷ Update parameters
8: $\quad cost \leftarrow \frac{1}{2n} \sum_{j=1}^{m} (h_j - y_j)^2$ ▷ Compute cost
9: $\quad$ **if** $|prev\_cost - cost| < \epsilon$ **then**
10: $\quad\quad$ **break** ▷ Convergence reached
11: $\quad$ **end if**
12: $\quad prev\_cost \leftarrow cost$
13: **end for**
14: **return** a

Example

The complexity order of this algorithm is $\mathcal{O}(n_{iter}(m+1)n)$, which is generally faster than the explicit determination whenever $n_{iter} < n$.
Let us work with the following example:

```
from sklearn import datasets
# Diabetes dataset
diabetes = datasets.load_diabetes()
```

Now, how we test if a variable is really relevant to explain the change in the dependent variable? We use the **t-test.**

T-test in Simple Linear Regression

Model and Hypotheses

- Simple Linear Regression Model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- Hypotheses:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

T-test in Simple Linear Regression I

- T-statistic:

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

- Standard Error:

$$SE(\hat{\beta}_1) = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2/(n-2)}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

- Degrees of Freedom: $df = n - 2$
- P-value (two-tailed test):

$$p\text{-value} = 2 \cdot P(T > |t|)$$

where $T \sim t(df)$

T-test in Simple Linear Regression II

- Confidence Interval ($(1-\alpha)$):

$$\hat{\beta}_1 \pm t_{\alpha/2,df} \cdot SE(\hat{\beta}_1)$$

- Decision Rule: Reject H_0 if $|t| > t_{\alpha/2,df}$ or if p-value $< \alpha$
- Interpretation:
  - Small p-value ($< \alpha$): Evidence against H_0
  - Large $|t|$: Stronger evidence of relationship
  - CI not including 0: Significant at α level

In Python I

```python
import statsmodels.api as sm
import numpy as np
import pandas as pd

# Assume you have your data in X (features)
# and y (target)
# X should be a 2D array or DataFrame,
# y should be a 1D array or Series

# Add a constant term to the features
X = sm.add_constant(X)
# Fit the model
model = sm.OLS(y, X).fit()
```

In Python II

```
# Print out the statistics
print(model.summary())
# Coefficients
coefficients = model.params
# Standard errors
std_errors = model.bse
# t-values
t_values = model.tvalues
# p-values
p_values = model.pvalues
# Confidence intervals
conf_int = model.conf_int()
```

Dimension I

We are generally interested in variables that can be observed. The most basic property of such variables is that they have a dimension.

Table: Dimensions of physical variables in the LMT system

Variable	Dimension	Variable	Dimension
Mass	M	Work	$M\ L^2\ T^{-2}$
Length	L	Pressure	$M\ L^{-1}\ T^{-2}$
Time	T	Power	$M\ L^2\ T^{-3}$
Frequency	T^{-1}	Angle	$M^0\ L^0\ T^0$
Velocity	$L\ T^{-1}$	Velocity of sound	$L\ T^{-1}$
Acceleration	$L\ T^{-2}$	Density	$M\ L^{-3}$
Force	$M\ L\ T^{-2}$	Dynamic viscosity	$M\ L^{-1}\ T^{-1}$
Energy	$M\ L^2\ T^{-2}$	Kinematic viscosity	$L^2\ T^{-1}$

Dimension II

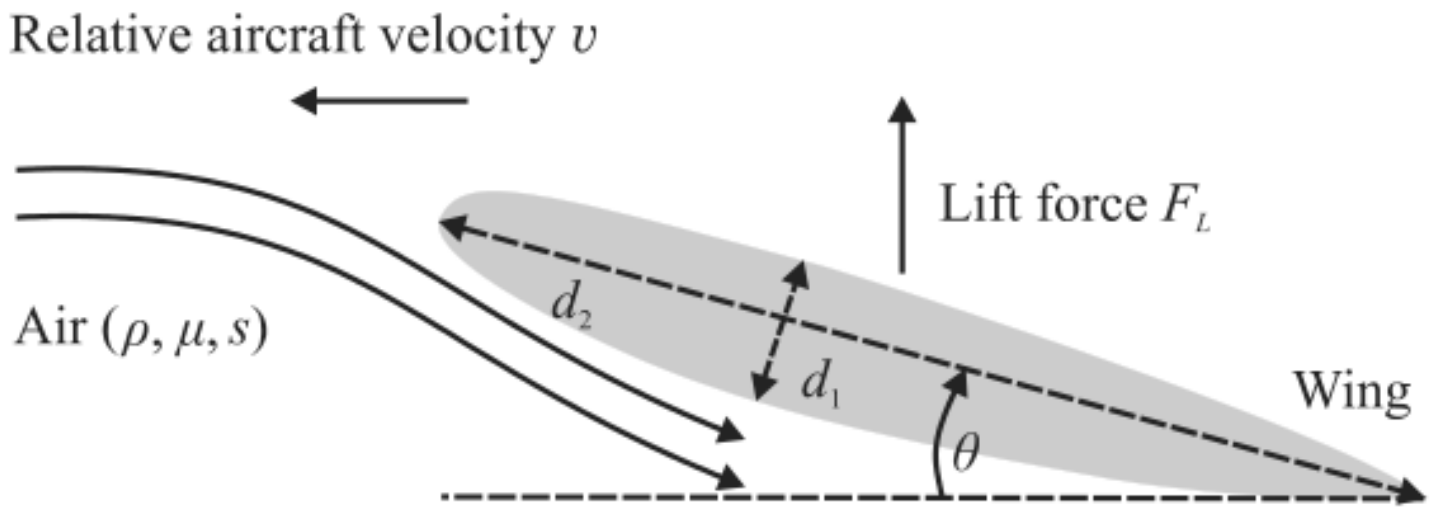

Figure: An illustration of factors that affect the lift force of aircrafts.

$$F_L = F_L(v, d_1, d_2, \theta, \rho, \mu, s).$$

$$F_L \sim \frac{ML}{T^2} \sim \frac{M}{L^3}\frac{L^2}{T^2}L^2 \sim \frac{C_L}{2}\rho v^2 d_2^2.$$

C_L is called lift coefficient.

Dimension III

$p = \rho v^2/2$ is called dynamic pressure, so that the previous relation can be written as

$$F_L = C_L p d_2^2.$$

Are the other variables irrelevant? **No**. They may be summarized in C_L, *i.e.*,

$$C_L = C_L(\theta, \frac{d_1}{d_2}, \frac{v}{s}, \dots).$$

Thus, a dimensionally correct equation for F_l can be written as a relation between nondimensional products,

$$\frac{F_L}{\rho v^2 d_2^2} = \frac{1}{2} C_L(\theta, \frac{d_1}{d_2}, \frac{v}{s}, \dots).$$

Dimension IV

Using the following abbreviations for the nondimensional products involved in the previous equation,

$$P_1 = \frac{F_L}{\rho v^2 d_2^2}, \quad P_2 = \theta, \quad P_3 = \frac{d_1}{d_2}, \quad P_4 = \frac{v}{s},$$

we get the following relation

$$P_1 = f(P_2, P_3, P_4, \dots).$$

Dimension V

Buckingham Theorem

If a physical process satisfies dimensional homogeneity and involves n physical variables that can be expressed in terms of k independent fundamental dimensions, then the process can be described by a set of $p = n - k$ dimensionless parameters $\Pi_1, \Pi_2, \ldots, \Pi_p$, such that:

$$f(\Pi_1, \Pi_2, \ldots, \Pi_p) = 0,$$

where f is some function of the p dimensionless parameters.

Dimensional Analysis I

Consider the following general product involving variables that may affect F_L,

$$F_L^a v^b d_1^c d_2^d \theta^e \rho^f \mu^g s^h.$$

$a, b, c, d, e, f, g, h \in \mathbb{R}$ are unknown. The conditions for having a nondimensional product is then

$$F_L^a v^b d_1^c d_2^d \theta^e \rho^f \mu^g s^h = c_1,$$

where c_1 is any constant that is independent of L, M and T. Thus,

$$[MLT^{-2}]^a [LT^{-1}]^b [L]^c [L]^d [L^0]^e [ML^{-3}]^f [ML^{-1}T^{-1}]^g [LT^{-1}]^h = c_2,$$

or equivalently,

$$M^{a+f+g} L^{a+b+c+d-3f-g+h} T^{-2a-b-g-h} = c_2.$$

Dimensional Analysis II

This leads to the following system of equations:

$$\begin{aligned} \text{For M:} \quad & a + f + g = 0 \\ \text{For L:} \quad & a + b + c + d - 3f - g + h = 0 \\ \text{For T:} \quad & -2a - b - g - h = 0 \end{aligned}$$

Let b, d and f be the dependent variables, so that a, c, e, g and h are the independent ones. Solving for the dependent ones gives

$$\begin{aligned} d &= -2a - c - g, \\ f &= -a - g, \\ b &= -2a - g - h. \end{aligned}$$

Dimensional Analysis III

The use of these expressions then provides

$$F_L^a v^{-2a-g-h} d_1^c d_2^{-2a-c-g} \theta^e \rho^{-a-g} \mu^g s^h = c_1,$$

or

$$\left(\frac{F_L}{\rho v^2 d_2^2}\right)^a \left(\frac{d_1}{d_2}\right)^c \theta^e \left(\frac{\mu}{\rho d_2 v}\right)^g \left(\frac{s}{v}\right)^h = c_1.$$

Five composite variables, as expected from Buckingham's Theorem. Of course, choosing different dependent variables when solving the linear relation will lead to different composite variables. Let us see further examples.

Reaction distance. Find a relation between the driver's reaction time T_R, the vehicular velocity v and the reaction distance D_R.

Dimensional Analysis IV

Braking Distance. Find a relation for the braking distance D_B having in mind that it is determined by the brake force F_B, the vehicular velocity v and the mass of the vehicle m.