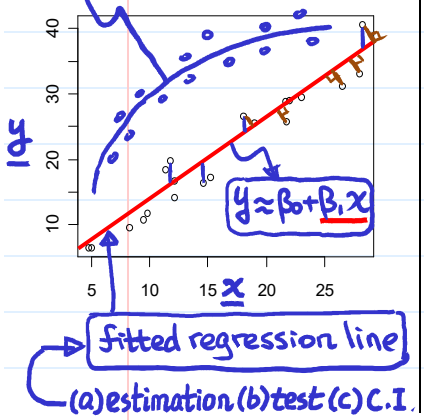
When to use regression analysis (linear model)?

What data? What objective?

- Regression: a statistical tool for investigating the "linearity relationship" between x and y.
  - ➢ causal relationship: examine the effects of x on y, i.e. how the changes in x *result in* the change in y (note that the statement implicitly assumes causal relationship), often seen in
    - DOE data (example?)
    - physical/chemical/engineering data (example?)



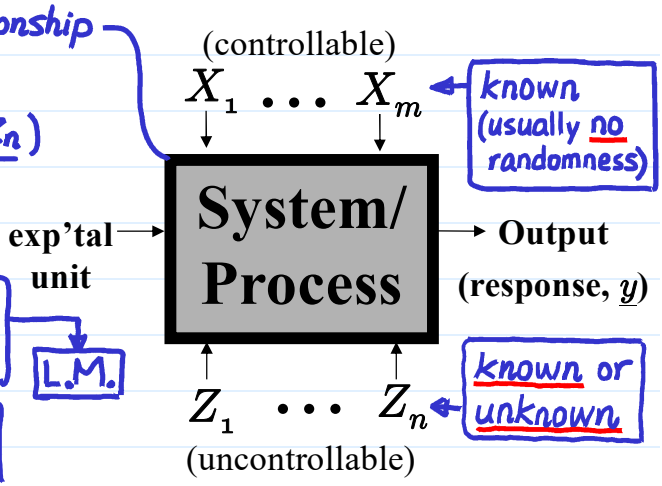
causal relationship: $y = f(x_1, \cdots, x_m, z_1, \cdots, z_n)$

unknown

approximate $f \approx \sum_{i=1}^{p} \beta_i g_i(x) + \varepsilon$

L.M.

use data to understand β_i's



- ➢ Even where no sensible causal relationship exists between x and y, we may wish to relate them by some sore of mathematical equation (rationale: sample from a multivariate normal population), often seen in
  - social data (example?)
  - economical data (example?)
  - medical data (example?)

X & Y: random variables (cf.)
(X, Y) ~ joint distribution
Y|X → linear model
(future lecture: sampling model)

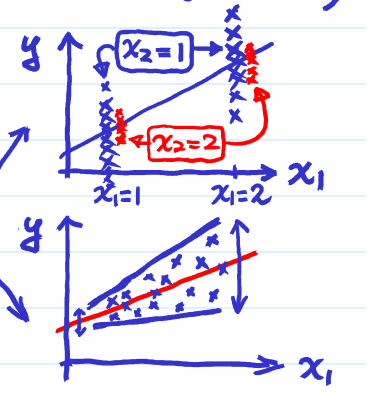
X usually not random in DOE

In some cases, what variable should be treated as the response & what should be the predictors could be a difficult problem → check LNp.2-8

- Q: Why develop the fitted line (fitted model)? Why it is useful?
  - ➢ numerical description of relationship → (y ↔ x, e.g. $\hat{\beta}_0 + \hat{\beta}_1 x$)
  - ➢ prediction → ($\hat{y}_x = ?$) ← (cf.) more difficult problem
  - ➢ coefficients may be meaningful → ($\hat{\beta}_i = ?$)
  - ➢ variation study → e.g. DOE, robust parameter designs
  - ➢ …

related to the mean function of y_x (i.e., $E(y_x)$)

$Var(y_x)$ is a function of x



- data type in regression analysis and some terminologies

A {*response*, *output*, *dependent*} variable Y is modeled or explained by p effects/functions of m {*predictor*, *input*, *independent*, *regressor*} variables $X_1,\ldots,X_m$

Recall continuous vs. discrete random variables

反應變數 解釋變數

Note. All "observed" data are discrete c.f.

- Y: "approximately" continuous c.f.
  - treated as random variables
  - 氣候 → 體溫 (normal, fever)
  - 溫度
  - 人數 → 學校人數 (1000, 20000); 中獎人數 (0, 1, 2, 3)
- $X_1,\ldots,X_m$: continuous and discrete (quantitative), categorical (qualitative) (Q: example?)
  - treated as fixed values
  - DOE
  - Y|X
  - with order
  - without order
- p=1, simple regression; p>1, multiple regression
  - 1 predictor, several effects
  - several predictors
  - # of effects
  - e.g. $y=\beta_0+\beta_1 x+\varepsilon$
- $X_1,\ldots,X_m$
  - all quantitative ⇒ multiple regression
  - quantitative+qualitative ⇒ analysis of covariance
  - all qualitative ⇒ analysis of variance (ANOVA)
  - multiple comparison
  - All ∈ Linear models
- more than one Y, multivariate regression
  - multivariate data analysis

- linear model:

$$\begin{aligned}\mathrm{E}(Y|X_1,\ldots,X_m) &= \textstyle\sum_{i=0}^{p}\beta_i\cdot g_i \\ \mathrm{Var}(Y|X_1,\ldots,X_m) &= \mathrm{Var}(\epsilon)=\sigma^2\end{aligned}$$

$$Y=\sum_{i=0}^{p}\beta_i\cdot g_i(X_1,\ldots,X_m)+\epsilon$$

$E(\epsilon)=0$, $\mathrm{Var}(\epsilon)=\sigma^2$

- 規律 (signal) → deterministic component → mean function
  - effects
  - & accurate (no error)
- 隨機 (noise) → error: random component → variance function

  - $X_1,\ldots,X_m$ are regarded as deterministic, i.e., no random phenomenon (when they are random variables, regard the linear model as conditional on $X_1,\ldots,X_m$)
    - how to determine g_i's?
    - ① conceptual ② empirical
    - $X_1,\cdots,X_m$ becomes fixed values
    - 效應
  - $g_0(X_1,\ldots,X_m)$, …, $g_p(X_1,\ldots,X_m)$: *known* functions of $X_1,\ldots,X_m$, called *effects*
  - (unknown) *parameters* β_0, …, β_p enter linearly
    - know the form of effects (g_i's), but don't know their magnitudes (β_i's)
  - variation due to random error only appears on y-axis

- Rationale: a general model for the relationship between Y and $X_1,\ldots,X_m$, is:

  $Y=f(X_1,\ldots,X_m)+\varepsilon$, where f is **unknown and arbitrary**

  - Why? consider the Taylor expansion of f → $f=\sum_{i=0}^{\infty}\beta_i x^i$

Note: # of parameters in f is infinite, usually do not have enough data to estimate f directly (globally), we have to assume that it has some more restricted form

  - c.f.
  - local approximation of f may be achievable by a linear model
    - with finite # of parameters
    - on the range of the observed $X_1,\cdots,X_m$ data

Note: Because the predictors can be transformed and combined in any way, linear models are actually very flexible.
- e.g. g_i's: x_1^t, x_1x_2, $e^{x_1+x_2}$, $\log\left(\frac{x_1}{x_2}\right)$, …

- Example:

Data

	Y	X_1	X_2	X_3
car1				
car2				

causality or association?

Y: fuel consumption of a particular model of car

X_1 : weight of the car (continuous? discrete? categorical?),

X_2 : horse power (continuous? discrete? categorical?),

X_3 : number of cylinders (continuous? discrete? categorical?)

X_4: 廠牌, Honda (1), Ford (2), Toyota (3), ··· (categorical)

➢ a general model: $Y = f(X_1, X_2, X_3) + \varepsilon$, where f is **unknown** & **arbitrary**

➢ possible statistical models ("local approximation" of f):

many

g_0 g_1 g_2 g_3

$Y = \beta_0 \underline{1} + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$ ⇒ a linear model

parameter

$Y = \beta_0 + \beta_1 X_1^2 + \beta_2 \log(X_2) + \beta_3 X_1 X_3 + \varepsilon$ ⇒ a linear model

$Y = \beta_0 + \beta_1 X_1^{\beta_2} + \varepsilon$ ⇒ not a linear model

a g_i?

$Y = \beta_1 X_1^{\beta_2} \varepsilon$ ⇒ not a linear model, but can be a linear model after taking log

$\log(Y) = \log(\beta_1) + \beta_2 \log(X_1) + \log(\varepsilon)$

"Y' "β1' "g" "ε'

❖ **Reading**: Faraway (2005, 1st ed.), 1.3, 2.1

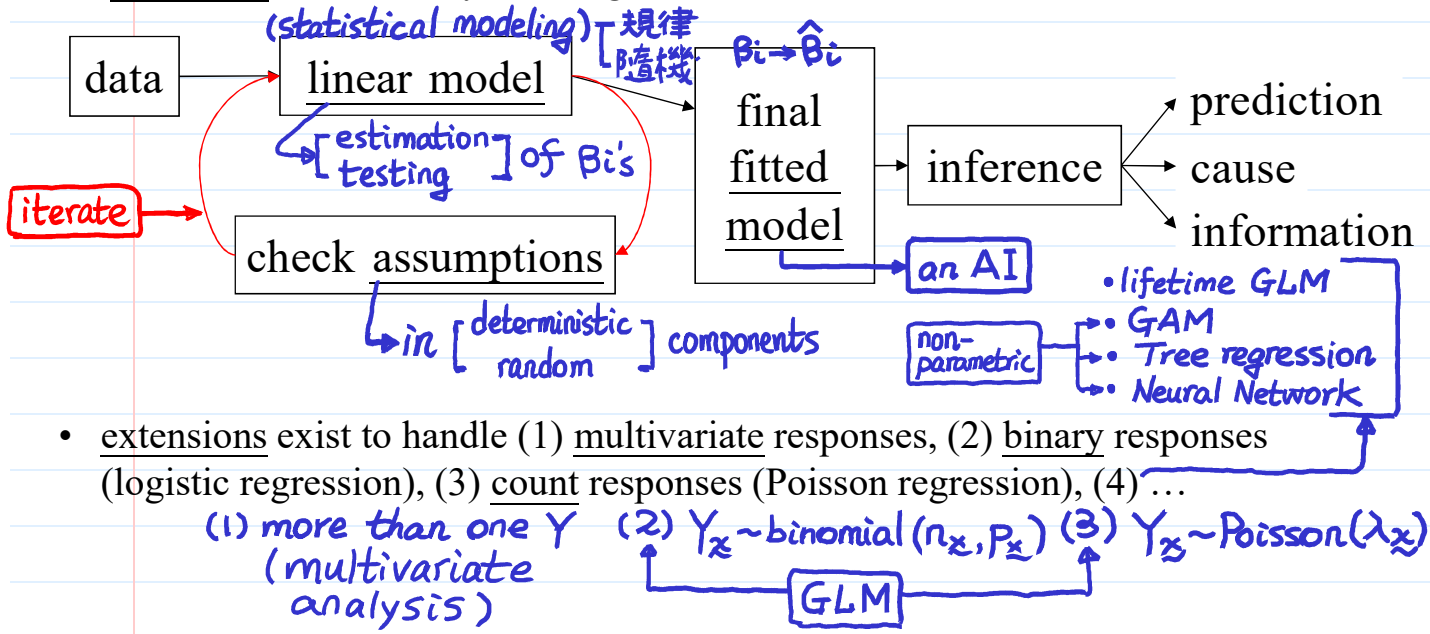
❖ **further reading**: D&S, 1.2 (meaning of linear model)

- some possible objectives of regression analysis

LNp.2-2

1. prediction of future observations
2. assessment of the effect of, or relationship between, explanatory variables on the response ← e.g. DOE in LNp.2-1
3. a general description of data structure ← e.g. association between some variables (LNp.2-2)
4. ...

- procedure of data analysis in regression



- extensions exist to handle (1) multivariate responses, (2) binary responses (logistic regression), (3) count responses (Poisson regression), (4) …

(1) more than one Y (multivariate analysis)

(2) $Y_x \sim$ binomial (n_x, p_x) (3) $Y_x \sim$ Poisson (λ_x)

GLM

History of regression

① linear structure ② data with random errors ③ parameter estimation

- Tobias Mayer (1750), made numerous observations to the libration of the moon with the purpose of determining the characteristics of the moon's orbit, and proposed the method of averages

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1}$$
$$Y^*_{p\times 1} = X^*_{p\times p}\beta_{p\times 1}$$
$$\Rightarrow \beta = (X^*)^{-1}Y^*$$

- Adrien Marie Legendre (1805), developed the method of least square

generalized → Gauss-Markov Thm (future lecture) ← future lecture

- Carl Friedrich Gauss (1809), claimed to have developed the method a few years earlier, and showed later least square is the best when the errors are normally distributed.

no need to get too pessimistic about mediocrity

- Sir Francis Galton (1875), coined the term "regression to mediocrity"
  - Family heights data

taller (shorter) parent ⇒ taller (shorter) child, but not as tall (short) as parent.

$$\left(\frac{y-\bar{y}}{\hat{\sigma}_y}\right) \overset{?}{=} 1\left(\frac{x-\bar{x}}{\hat{\sigma}_x}\right)$$

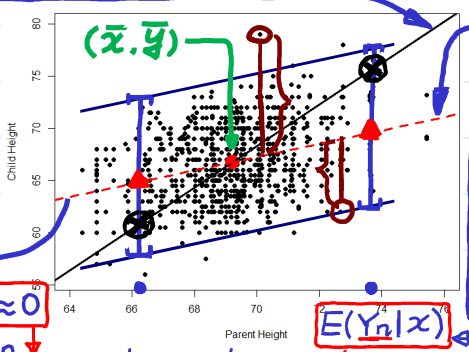
Standard score

like from 1 to 10

$$\left(\frac{y-\bar{y}}{\hat{\sigma}_y}\right) = \hat{\rho}\left(\frac{x-\bar{x}}{\hat{\sigma}_x}\right)$$

$\hat{\rho} < 1$

mean, least square line



need to consider the influence of ε

$E(Y_n|x)$

$Cor(X, Y_n) \approx 0$

$\hat{\rho}^n \approx 0$ when n large

$\Rightarrow \left(\frac{y_n-\bar{y}_n}{\hat{\sigma}_{y_n}}\right) \approx 0 \Rightarrow y_n \approx \bar{y}_n$

So, $\hat{\sigma}_{y_n} \approx 0$? No.

zero $\left(\frac{x-\bar{x}}{\hat{\sigma}_x}\right)$ → 1st $\hat{\rho}\left(\frac{x-\bar{x}}{\hat{\sigma}_x}\right) = \left(\frac{y_1-\bar{y}_1}{\hat{\sigma}_{y_1}}\right)$ → 2nd $\hat{\rho}^2\left(\frac{x-\bar{x}}{\hat{\sigma}_x}\right) = \left(\frac{y_2-\bar{y}_2}{\hat{\sigma}_{y_2}}\right)$ → ⋯ → nth $\hat{\rho}^n\left(\frac{x-\bar{x}}{\hat{\sigma}_x}\right) = \left(\frac{y_n-\bar{y}_n}{\hat{\sigma}_{y_n}}\right)$

  - This so called regression effect is not a natural law. It is actually a mathematical result.

In many applications, regression effect is not of interest. (e.g., different types of variables)

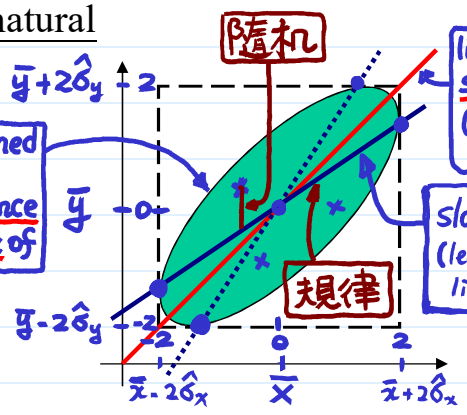
$\hat{\rho}$: correlation coefficient of (X,Y)

$$\left(\frac{y-\bar{y}}{\hat{\sigma}_y}\right) = \hat{\rho}\left(\frac{x-\bar{x}}{\hat{\sigma}_x}\right)$$

$-1 < \hat{\rho} < 1$

$$\left(\frac{x-\bar{x}}{\hat{\sigma}_x}\right) = \hat{\rho}\left(\frac{y-\bar{y}}{\hat{\sigma}_y}\right)$$

determined by the covariance matrix of (X,Y), LNp.4-5



line with slope = 1 (1st principal component)

slope $\hat{\rho} < 1$ (least square line)

  - Misinterpretation (Note. Regression effect does not imply causality.)

check LNp.2-2

- IQ of a couple (husband, wife) — h: response, w: predictor; w: response, h: predictor
- Ineffective treatment, measurements before and after treatment — before: predictor; after: response
- Punishment works and reward does not.

Should use 2-sample comparison or paired comparison

  - Warm and Cold — pro / con
    - Family height data
    - Josef Stalin: "A single death is a tragedy; a million deaths is a statistic!" (一個人的死是一場悲劇，百萬個人的死只是一個統計數據！)

❖ **Reading**: Faraway (2005, 1st ed.), 1.4

❖ **Further reading**: D&S, 1.8;
Cathy O'Nell (2016), Weapons of Math Destruction (中譯：大數據的傲慢與偏見).