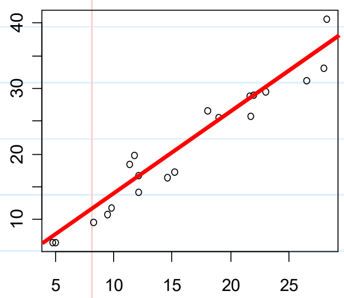
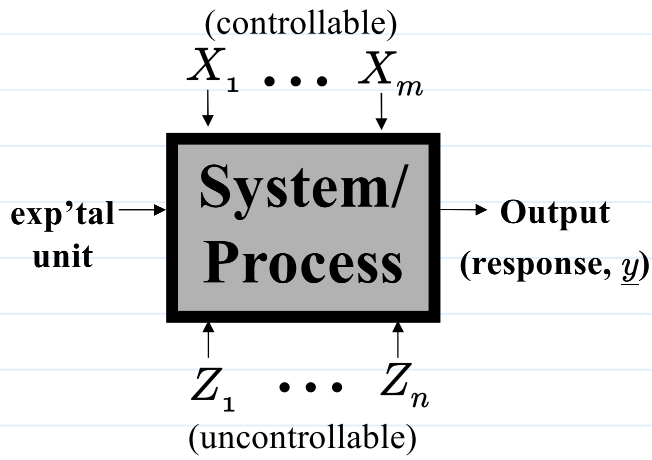
When to use regression analysis (linear model)?

- 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?)

Example





  - ➢ 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?)

- **Q**: Why develop the fitted line (fitted model)? Why it is useful?
  - ➢ numerical description of relationship
  - ➢ prediction
  - ➢ coefficients may be meaningful
  - ➢ variation study
  - ➢ …

- 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 $X_1,\ldots,X_m$

➢ Y: "approximately" continuous

➢ $X_1,\ldots,X_m$: continuous and discrete (quantitative), categorical (qualitative) (**Q**: example?)

➢ p=1, simple regression; p>1, multiple regression

➢ $X_1,\ldots,X_m$

all quantitative	$\Rightarrow$	multiple regression
quantitative+qualitative	$\Rightarrow$	analysis of covariance
all qualitative	$\Rightarrow$	analysis of variance (ANOVA)

➢ more than one Y, multivariate regression

- linear model:

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

$$Y=\underbrace{\sum_{i=0}^{p}\beta_i\cdot g_i(X_1,\ldots,X_m)}_{\text{deterministic component}\;\to\;\text{mean function}}\;+\;\underbrace{\epsilon}_{\text{error: random component}\;\to\;\text{variance function}}$$

➢ $X_1,\ldots,X_m$ are regarded as deterministic, i.e., no random phenomenon (when they are random variables, regard the linear model as conditional on $X_1,\ldots,X_m$)

➢ $g_0(X_1,\ldots,X_m)$, ..., $g_p(X_1,\ldots,X_m)$: *known* functions of $X_1,\ldots,X_m$, called *effects*

➢ (unknown) *parameters* β_0, ..., β_p enter linearly

➢ 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:

$$Y=f(X_1,\ldots,X_m)+\varepsilon,\ \text{ where } f \text{ is } \textbf{unknown and arbitrary}$$

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

➢ local approximation of f may be achievable by a linear model

- **Note:** Because the predictors can be transformed and combined in any way, linear models are actually very flexible.

- Example:

  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?)

  - a general model: $Y = f(X_1, X_2, X_3) + \varepsilon$, where f is **unknown** & **arbitrary**
  - possible statistical models ("local approximation" of f):

$$Y = \beta_0\, 1 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \quad \Rightarrow \text{a linear model}$$

$$Y = \beta_0 + \beta_1 X_1^2 + \beta_2 \log(X_2) + \beta_3 X_1 X_3 + \varepsilon \quad \Rightarrow \text{a linear model}$$

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

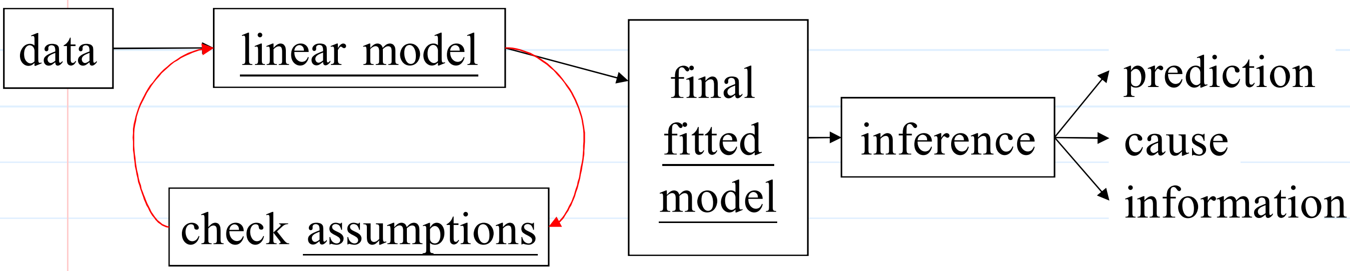
$$Y = \beta_1 X_1^{\beta_2} \varepsilon$$

$\Rightarrow$ not a linear model, but can be a linear model after taking log

❖ **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
  1. prediction of future observations
  2. assessment of the effect of, or relationship between, explanatory variables on the response
  3. a general description of data structure
  4. ...

- procedure of data analysis in regression



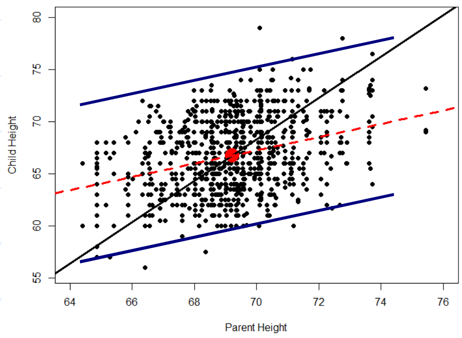
- extensions exist to handle (1) multivariate responses, (2) binary responses (logistic regression), (3) count responses (Poisson regression), (4) …

History of regression

- 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
- Adrien Marie Legendre (1805), developed the method of least square
- 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.
- Sir Francis Galton (1875), coined the term "regression to mediocrity"
  - ➢ Family heights data

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

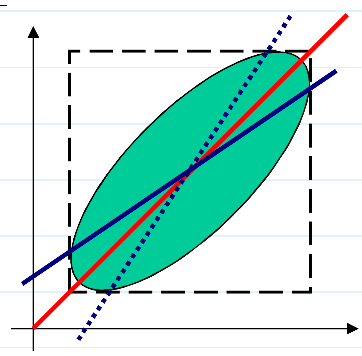
$$\left(\frac{y-\bar{y}}{\hat{\sigma}_y}\right) = \hat{\rho}\left(\frac{x-\bar{x}}{\hat{\sigma}_x}\right)$$



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

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

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



➢ Misinterpretation (Note. Regression effect does not imply causality.)
- IQ of a couple (husband, wife)
- Ineffective treatment, measurements before and after treatment
- Punishment works and reward does not.

➢ Warm and Cold
- 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).

Matrix representation

- Given the data matrix,

Y	X_1	X_2	...	X_m
y_1	x_{11}	x_{12}	...	x_{1m}
y_2	x_{21}	x_{22}	...	x_{2m}
...		...		
y_n	x_{n1}	x_{n2}	...	x_{nm}

a row: one group of observations

a column: one variable (response or predictor)

- We may write a linear model as follows (*functional form*): for $i = 1, 2, ..., n$,

$$y_i = \beta_0 + \beta_1\, g_1(x_{i1},...,x_{im}) + \beta_2\, g_2(x_{i1},...,x_{im}) + ... + \beta_{p-1}\, g_{p-1}(x_{i1},...,x_{im}) + \varepsilon_i,$$

Y	**1**	g_1	g_2	...	g_{p-1}
y_1	1	g_{11}	g_{12}	...	g_{1p-1}
y_2	1	g_{21}	g_{22}	...	g_{2p-1}
...			...		
y_n	1	g_{n1}	g_{n2}	...	g_{np-1}

a row: one group of observations

a column: response or effect

where $g_{ij} = g_j(x_{i1},...,x_{im})$

- the expression is (i) ugly notation (ii) conceptually awkward
- matrix/vector notation is more elegant

Y	=	**1**	+	g_1	+	g_2	+	...	+	g_{p-1}	+	ε
y_1	=	1	+	g_{11}	+	g_{12}	+	...	+	g_{1p-1}	+	ε_1
y_2	=	1	+	g_{21}	+	g_{22}	+	...	+	g_{2p-1}	+	ε_2
...	=					...					+	...
y_n	=	1	+	g_{n1}	+	g_{n2}	+	...	+	g_{np-1}	+	ε_n

a row: one group of observations

a column: response or effect

- Matrix form of the linear model:

$$Y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & g_{11} & \cdots & g_{1p-1} \\ 1 & g_{21} & \cdots & g_{2p-1} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & g_{n1} & \cdots & g_{np-1} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_{p-1} \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdots \\ \epsilon_n \end{bmatrix}.$$

and $E(\boldsymbol{\varepsilon}) = \boldsymbol{0}$ and $var(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$ (**Note**: the assumption that errors are normally distributed is not required at the estimation stage)

- Example 1 (no predictor model, seen in one sample problem):

y_i's are i.i.d. with mean μ and variance σ^2, i=1,...,n

$$Y=\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix},\ X=\begin{bmatrix} 1 \\ 1 \\ \dots \\ 1 \end{bmatrix},\ \boldsymbol{\beta}=\begin{bmatrix} \mu \end{bmatrix},\ \boldsymbol{\varepsilon}=\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \dots \\ \epsilon_n \end{bmatrix}.$$

- Example 2 (the model in two sample problem):

z_i's are i.i.d. with mean μ_1 and variance σ^2, i=1,...,m
w_j's are i.i.d. with mean μ_2 and variance σ^2, j=1,...,n

$$Y=\begin{bmatrix} z_1 \\ \dots \\ z_m \\ w_1 \\ \dots \\ w_n \end{bmatrix},\ X=\begin{bmatrix} 1 & 0 \\ \dots & \dots \\ 1 & 0 \\ 0 & 1 \\ \dots & \dots \\ 0 & 1 \end{bmatrix},\ \boldsymbol{\beta}=\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix},\ \boldsymbol{\varepsilon}=\begin{bmatrix} \epsilon_1 \\ \dots \\ \epsilon_m \\ \delta_1 \\ \dots \\ \delta_n \end{bmatrix}.$$

Y	X	$g_1(X)$	$g_2(X)$
z_1	1	1	0
...	...	...	...
z_m	1	1	0
w_1	2	0	1
...	...	...	...
w_n	2	0	1

❖ **Reading**: Faraway (2005, 1st ed.), 2.2
❖ **Further reading**: D&S, 4.1