MULTIVARIATE GAUSSIAN DISTRIBUTION

24/03

$\underline{X}$ random vector, $\underline{X} \in \mathbb{R}^p$

$\mathcal{J}_{\underline{X}} : \mathcal{B}(\mathbb{R}^p) \to [0,1]$ — the distribution characterizes $\underline{X}$

($\mathcal{B}(\mathbb{R}^p)$: Borel sets of $\mathbb{R}^p$)

$\mathcal{J}_{\underline{X}}(B) = \mathbb{P}(\underline{X} \in B) \quad \forall B \in \mathcal{B}(\mathbb{R}^p)$

$\mathcal{J}_{\underline{X}}(B) = \int_B f(\underline{x})\, d\underline{x}$ $:=$ DENSITY: 1. $f \geq 0$ 2. $\int_{\mathbb{R}^p} f(\underline{x})\, d\underline{x} = 1$

Def. $\underline{X} \in \mathbb{R}^p$ random vector with density f. Let $\underline{\mu} \in \mathbb{R}^p$, Σ $p \times p$ pos. def.. Then:

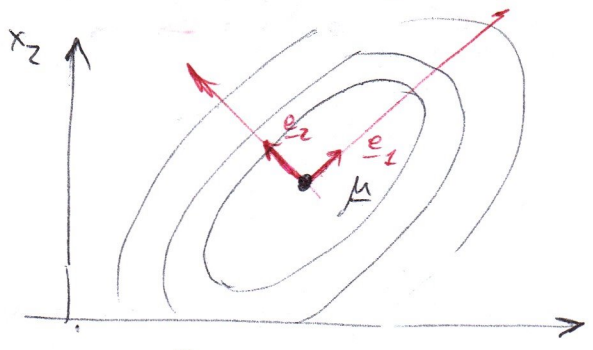
$\underline{X} \sim N_p(\underline{\mu}, \Sigma)$ if $f(\underline{x}) = \frac{1}{\sqrt{(2\pi)^p \det(\Sigma)}} \exp^{-\frac{1}{2}(\underline{x}-\underline{\mu})^T \Sigma^{-1} (\underline{x}-\underline{\mu})}$

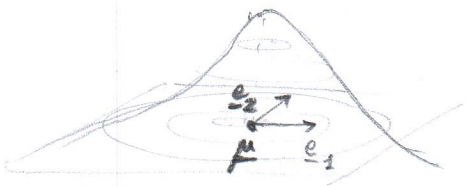
$\propto \exp\left(-\frac{1}{2} d^2_{\Sigma^{-1}}(\underline{x}, \underline{\mu})\right)$ (Mahalanobis's distance)

CONTOUR PLOTS FOR f: (lines that identify points where the elevation of the density is the same)

$\{\underline{x} \in \mathbb{R}^p : f(\underline{x}) = \text{constant}\} = \{\underline{x} \in \mathbb{R}^p : (\underline{x}-\underline{\mu})^T \Sigma^{-1} (\underline{x}-\underline{\mu}) = \text{const}^2\}$

$= \{\underline{x} \in \mathbb{R}^p : d_{\Sigma^{-1}}(\underline{x}, \underline{\mu}) = \text{const}^2\}$





if $\Sigma = \sum_{i=1}^p \lambda_i \underline{e}_i \underline{e}_i^T$ $\Longrightarrow$ the axes of the ellipse are the eigenvectors

$\Sigma^{-1} = \sum_{i=1}^p \frac{1}{\lambda_i} \underline{e}_i \underline{e}_i^T$

directions identified by the PCs, so with the gaussian distribution the PCA identifies the axes of the ellipses that are the contour plots for the density

$\Longrightarrow$ $\begin{cases} \underline{e}_1, \dots, \underline{e}_p & : \text{axes} \\ \sqrt{\lambda_1}, \dots, \sqrt{\lambda_p} & : \text{length} \end{cases}$

Prop. $\underline{X} \sim N_p(\underline{\mu}, \Sigma) \to \begin{cases} \mathbb{E}[\underline{X}] = \underline{\mu} \\ \text{Cov}(\underline{X}) = \Sigma \end{cases}$

Theorem: (Characterization) $\underline{X} \sim N_p(\underline{\mu}, \Sigma) \iff \underline{a}^T \underline{X} \sim N_1(\underline{a}^T \underline{\mu}, \underline{a}^T \Sigma \underline{a}) \quad \forall \underline{a} \in \mathbb{R}^p$

! If we have a gaussian vector $\Rightarrow$ $\forall$ linear comb. of the components it gaussian
If $\forall$ lin. comb. of the components of $\underline{X}$ is gaussian $\Longrightarrow$ $\underline{X}$ is gaussian

proof. Use characteristic functions / moment generating function. (■)

Impact of the theorem < practical / theoretical : What do we have to do when we want to check gaussianity for a multivariate dataset (we project the data (random vector) on linear spaces and then check if the projections are gaussian (we know how to check gaussianity in 1D : Shapiro,..)

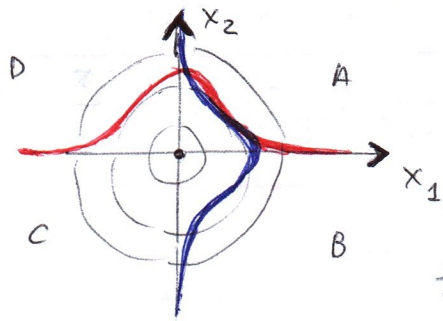
Corollary: If $\underline{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}$, $\underline{X} \sim N_p(\underline{\mu}, \Sigma) \Rightarrow X_i \sim N_1(\mu_i, \sigma_{ii})$ $\Sigma = [\sigma_{ij}]$

($\not\Leftarrow$)

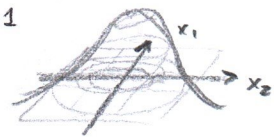
"proof."

$X_i = \underline{u}_i^T \underline{X}$: $\underline{u}_i = [0, 0, \dots, 0, \underset{i}{1}, 0, \dots, 0]^T$ (by thm.) ■

Ex. ($\not\Leftarrow$) : $\underline{X} \sim N_2(\underline{0}, I)$

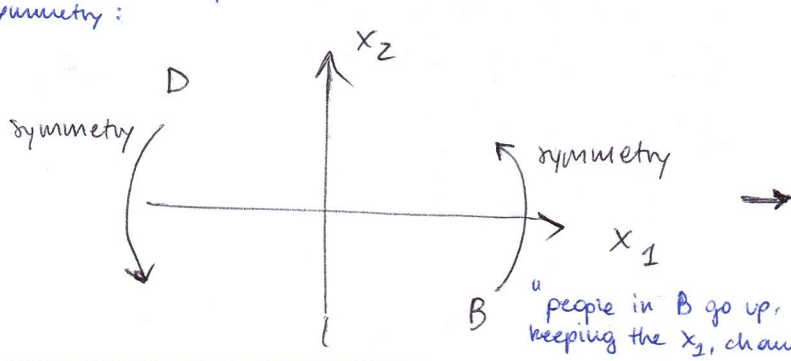


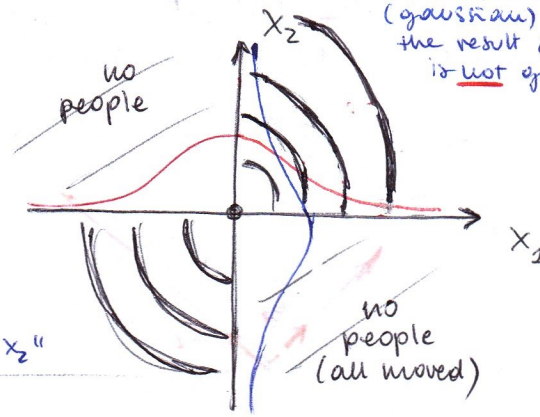
= } singular densities (X_1, X_2)
(marginal densities of X_1 and X_2)



people are distributed in A and C with the same contour lines. ∴ the marginal distr. (among X_1 and X_2) are still the same (gaussian) but the result (in 3D) is not gaussian

now we move by symmetry :





Prop. $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$, A $q \times p$ matrix

$\Rightarrow A\underline{X} \in \mathbb{R}^q$, $\boxed{A\underline{X} \sim N_q(A\underline{\mu}, A\Sigma A^T)}$ (Gaussianity is preserved!)

proof.

let $\underline{a} \in \mathbb{R}^q$ need $\underline{a}^T(A\underline{X})$ has a Gaussian dist.

$\underline{a}^T(A\underline{X}) = (\underline{a}^T A)\underline{X} = (A^T\underline{a})^T \underline{X}$ ($A^T\underline{a} \in \mathbb{R}^p \Rightarrow$ Thm. (charact.))

thm. $\Rightarrow (A^T\underline{a})^T \underline{X} \sim N_1((A^T\underline{a})^T\underline{\mu}, (A^T\underline{a})^T \Sigma (A^T\underline{a}))$

$\sim N_1(\underline{a}^T(A\underline{\mu}), \underline{a}^T A\Sigma A^T \underline{a})$ $\forall \underline{a} \in \mathbb{R}^q$

thm. $\Rightarrow A\underline{X} \sim N_q(A\underline{\mu}, A\Sigma A^T)$ ■

Prop. $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$, $\underline{d} \in \mathbb{R}^p \Rightarrow \boxed{\underline{X} + \underline{d} \sim N_p(\underline{\mu} + \underline{d}, \Sigma)}$

proof. for us. ■

(using again the characterization)

Notations

$$\underline{X} = \begin{bmatrix} \underline{X}_1 \\ \underline{X}_2 \end{bmatrix}, \quad \underline{X}_1 \in \mathbb{R}^q, \quad \underline{X}_2 \in \mathbb{R}^{p-q} \qquad q < p$$

$$\underline{\mu} = \begin{bmatrix} \underline{\mu}_1 \\ \underline{\mu}_2 \end{bmatrix}, \quad \underline{\mu}_1 \in \mathbb{R}^q, \quad \underline{\mu}_2 \in \mathbb{R}^{p-q}$$

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \quad \text{s.t.} \quad \begin{cases} \Sigma_{11} \in \mathbb{R}^{q \times q} \\ \Sigma_{12} \in \mathbb{R}^{q \times (p-q)} \\ \Sigma_{21} \in \mathbb{R}^{(p-q) \times q} \\ \Sigma_{22} \in \mathbb{R}^{(p-q) \times (p-q)} \end{cases}$$

! Not only the singular components have the gaussian distribution, any subset of the components have gaussian distribution

Prop. $\underline{X} = \begin{bmatrix} \underline{X}_1 \\ \underline{X}_2 \end{bmatrix} \sim N_p(\mu, \Sigma) \Longrightarrow \underline{X}_1 \sim N_q(\mu_1, \Sigma_{11})$

$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \sim N \Rightarrow \begin{bmatrix} X_1 \\ X_3 \end{bmatrix} \sim N$

proof.

let $A = [I \quad 0]$ $\quad I \in \mathbb{R}^{q \times q}$, $0 \in \mathbb{R}^{q \times (p-q)}$

$\Longrightarrow A\underline{X} \sim N_q(A\mu, A\Sigma A^T)$:

$A\underline{X} = \underline{X}_1$

$A\mu = \mu_1$

$A\Sigma A^T = \Sigma_{11}$ ∎

Theorem: $\underline{X} = \begin{bmatrix} \underline{X}_1 \\ \underline{X}_2 \end{bmatrix} \sim N_p\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$:

$$\underline{X}_1 \perp\!\!\!\perp \underline{X}_2 \iff \Sigma_{12} = 0 \quad (\equiv \Sigma_{21} = 0)$$

this is not true in a generic case

proof. $f_{\underline{X}} = f_{\underline{X}_1} \cdot f_{\underline{X}_2}$ (..., for us) ∎

Theorem: $X = \begin{bmatrix} \underline{X}_1 \\ \underline{X}_2 \end{bmatrix} \sim N_p\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right) \qquad (\det(\Sigma_{22}) \neq 0)$

$$\Longrightarrow \underline{X}_1 | \underline{X}_2 = x_2 \sim N_q\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\underline{x}_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$$

conditional distribution of a subset of components of $\underline{X}$ once we know the other part of the vector

proof.

$\underline{Z} \sim N_p(\underline{0}, I)$

$\underline{Z} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_p \end{bmatrix}$, $\mathrm{cov}(\underline{Z}) = \begin{bmatrix} 1 & & \emptyset \\ & \ddots & \\ \emptyset & & 1 \end{bmatrix}$

$\iff z_1, \dots, z_p \overset{iid}{\sim} N(0,1)$

$A \in \mathbb{R}^{p \times p}$: $A\underline{Z} \sim N_p(\underline{0}, AIA^T) \sim N_p(\underline{0}, AA^T)$

$A\underline{Z} + \mu \sim N_p(\mu, AA^T)$

let $A = \Sigma^{\frac{1}{2}} = \sum_{i=1}^{p} \sqrt{\lambda_i}\, \underline{e}_i \cdot \underline{e}_i^T$

$\Sigma^{\frac{1}{2}}\underline{Z} + \mu \sim N_p(\mu, \Sigma)$

$\underline{X} \sim N_p(\mu, \Sigma)$

$\Sigma^{-\frac{1}{2}}(\underline{X} - \mu) \sim N_p(\underline{0}, I)$

We're generating any multivariate gaussian distr. starting from gaussian distr. whose components are iid with standard normal distribution in $\mathbb{R}$ simply by linear trasformations

proof.

$$A = \begin{bmatrix} I\ (\in \mathbb{R}^{q\times q}) & -\Sigma_{12}\Sigma_{22}^{-1}\ (\in \mathbb{R}^{q\times(p-q)}) \\ 0\ (\in \mathbb{R}^{(p-q)\times q}) & I\ (\in \mathbb{R}^{(p-q)\times(p-q)}) \end{bmatrix}$$

$$A(\underline{X}-\underline{\mu}) = A\begin{bmatrix} \underline{X}_1-\underline{\mu}_1 \\ \underline{X}_2-\underline{\mu}_2 \end{bmatrix} = \begin{bmatrix} \underline{X}_1-\underline{\mu}_1-\Sigma_{12}\Sigma_{22}^{-1}(\underline{X}_2-\underline{\mu}_2) \\ \underline{X}_2-\underline{\mu}_2 \end{bmatrix} \sim N_p\left(\begin{bmatrix} \underline{0} \\ \underline{0} \end{bmatrix}, A\Sigma A^T\right)$$

$$\sim N_p\left(\begin{bmatrix} \underline{0} \\ \underline{0} \end{bmatrix}, \begin{bmatrix} \Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0 \\ 0 & \Sigma_{22} \end{bmatrix}\right)$$

by simple linear transf. we can prove $\perp\!\!\!\perp$ between two functions of the original vector

$$\underline{X}_1-\underline{\mu}_1-\Sigma_{12}\Sigma_{22}^{-1}(\underline{X}_2-\underline{\mu}_2) \perp\!\!\!\perp \underline{X}_2-\underline{\mu}_2$$

$$\underbrace{\underline{X}_1-\underline{\mu}_1-\Sigma_{12}\Sigma_{22}^{-1}(\underline{X}_2-\underline{\mu}_2)}_{:=\underline{W}} \sim N_q(\underline{0}, \Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

$\underline{W} \mid \underline{X}_2 = \underline{x}_2$?

$$\underline{W}\mid \underline{X}_2=\underline{x}_2 \sim N_q(\underline{0}, \Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

(independence)

$$\underline{W}\mid \underline{X}_2=\underline{x}_2 = \underline{X}_1-\underline{\mu}_1-\Sigma_{12}\Sigma_{22}^{-1}(\underline{x}_2-\underline{\mu}_2)$$

Given $\underline{X}_2=\underline{x}_2$ $\Rightarrow$ $\underline{X}_1-\underline{\mu}_1-\Sigma_{12}\Sigma_{22}^{-1}(\underline{x}_2-\underline{\mu}_2) \sim N_q(\underline{0}, \Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$

by transl. $\Rightarrow$ $\underline{X}_1\mid \underline{X}_2=\underline{x}_2 \sim N_q(\mu_1+\Sigma_{12}\Sigma_{22}^{-1}(\underline{x}_2-\underline{\mu}_2), \Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$ ■

$$\boxed{Cov(\underline{X}_1\mid \underline{X}_2=\underline{x}_2) = \Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}$$

does not depend on $\underline{x}_2$

$:=$ PARTIAL COVARIANCES.

! $\Rightarrow$ even without knowing the value of $\underline{x}_2$ we already know how the information will be modified (how the uncertainty will be modified in inference for $\underline{X}_1$

Example: $p=2$: $\underline{X} = \begin{bmatrix} X \\ Y \end{bmatrix} \sim N_2\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_{XX} & \sigma_{XY} \\ \sigma_{YX} & \sigma_{YY} \end{bmatrix}\right)$

$\boxed{Y} \sim N(\mu_Y, \sigma_{YY})$

$\mathbb{P}(Y \in [\mu_Y \pm 2\sqrt{\sigma_{YY}}]) = 0.95$

$\boxed{Y\mid X=x} \sim N_1(\mu_Y+\sigma_{XY}\sigma_{XX}^{-1}(x-\mu_X), \sigma_{YY}-\sigma_{XY}\sigma_{XX}^{-1}\sigma_{XY})$

$\sim N_1\left(\mu_Y+\frac{\sigma_{XY}}{\sigma_{XX}}(x-\mu_X), \sigma_{YY}-\frac{\sigma_{XY}^2}{\sigma_{XX}}\right)$

$\rho_{XY} = \frac{\sigma_{XY}}{\sqrt{\sigma_{XX}\sigma_{YY}}}$ $[-1,1]$

$\Rightarrow$ $\boxed{Y\mid X=x} \sim N_1\left(\mu_1+\frac{\sigma_{XY}}{\sigma_{XX}}(x-\mu_X), \sigma_{YY}\left(1-\frac{\sigma_{XY}^2}{\sigma_{XX}\sigma_{YY}}\right)\right)$

$\sim N_1\left(\mu_1+\frac{\sigma_{XY}}{\sigma_{XX}}(x-\mu_X), \sigma_{YY}(1-\rho^2)\right)$

$$Y \sim N(\mu_Y, \sigma_{YY})$$
$$Y|X=x \sim N(\cdots, \sigma_{YY}(1-\rho^2))$$

after the information (if $\rho \neq 0$) the variability of y decreases

$$E[Y|X=x] = \mu_Y + \frac{\sigma_{XY}}{\sigma_{XX}}(x-\mu_X) = \text{function}(x) := y$$

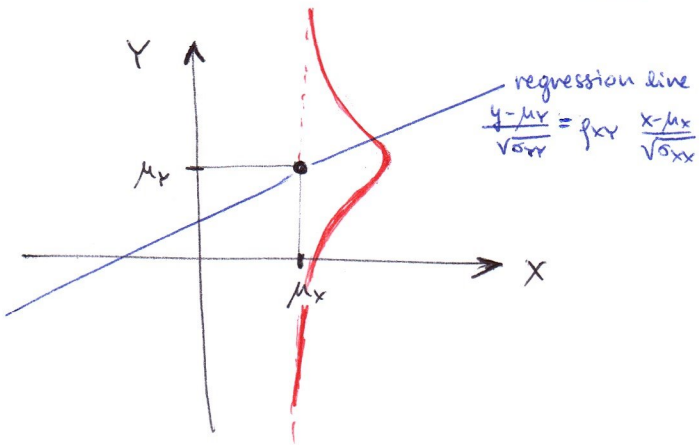
$$y = \mu_Y + \frac{\sigma_{XY}}{\sigma_{XX}}(x-\mu_X) = \text{linear regression function}$$

$$\frac{y-\mu_Y}{\sqrt{\sigma_{YY}}} = \frac{\sigma_{XY}}{\sqrt{\sigma_{XX}}\sqrt{\sigma_{YY}}} \cdot \frac{x-\mu_X}{\sqrt{\sigma_{XX}}}$$

$:= \rho_{XY}$ = CORRELATION COEFFICIENT

$$\frac{y-\mu_Y}{\sqrt{\sigma_{YY}}} = \rho_{XY} \frac{x-\mu_X}{\sqrt{\sigma_{XX}}}$$

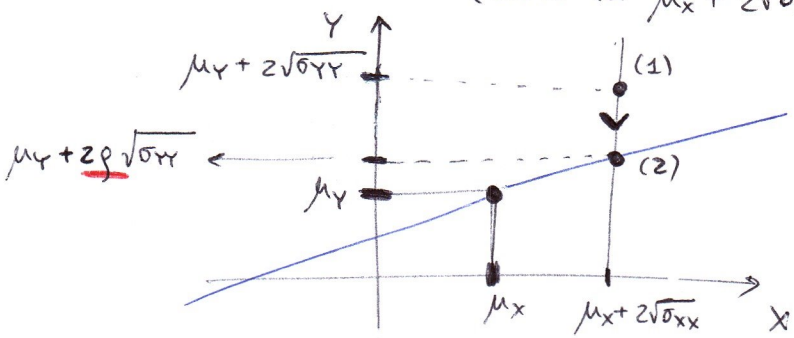
stol. dlate for Y — Stol. dlate for X



If we don't have any information and we want to predict y, the best prediction will be centered on μ_Y and will have some sort of gaussian distribution (—)

But suppose that we have some informations: mean of x (we're in $\mu_X + 2\sqrt{\sigma_{XX}}$ line).

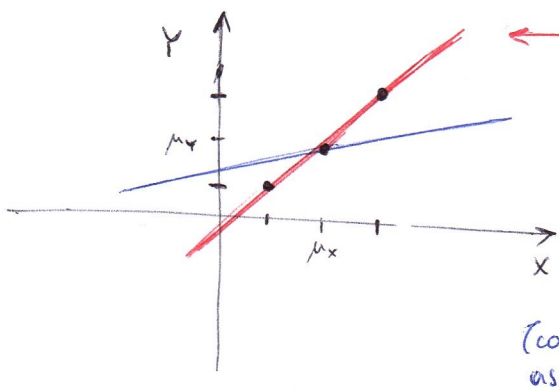
Suppose we're far from the



Can we predict y?

We may think that since it's $2\sqrt{\sigma_{XX}}$ above the mean maybe it will be $2\sqrt{\sigma_{YY}}$ above μ_Y, so maybe we're in (1).

What the regression line is telling us is that's not the, the best prediction is (2)



← this is what we thought it would be like (since it's two d standard deviations from μ_X then it'll be two standard deviation from μ_Y) but it's not like it

REGRESSION EFFECT ($|\rho|<1$)

(! Attention: REGRESSION FALLACY)

(consist in trying to interpret the regression effect as if there ∃ a cause-effect)

Statistical Methods in Medical Research 1997; **6**: 103–114

Regression towards the mean, historically considered

Stephen M Stigler Department of Statistics, University of Chicago, Chicago, Illinois, USA

Regression. It is a universal rule that the unknown kinsman in any degree of any specified man, is probably more mediocre than he. (Francis Galton, 1886)

The simple yet subtle concept of regression towards the mean is reviewed historically. Verbal, geometric, and mathematical expressions of the concept date to the discoverer of the concept, Francis Galton. That discovery and subsequent understanding (and misunderstanding) of the concept are surveyed.

1 Introduction

Regression towards the mean is an elementary concept in statistics. When properly understood, it is transparent to the point of being obvious. Yet despite its simplicity, it has been consistently misunderstood and it has repeatedly been the source of major errors in analysis, some with significant policy implications, attracting such names as 'the regression paradox', 'the regression fallacy' and 'the regression trap'. Milton Friedman has written 'I suspect that the regression fallacy is the most common fallacy in the statistical analysis of economic data',[1] a sentiment that could with justice be carried over to any other field where multivariate data are employed for the analysis and formulation of policies.

To understand the nature of this phenomenon, of how a simple idea could cause so much difficulty, it will be useful to examine the history of the idea, because the historical origins reveal a number of ways of interpreting it that could, if more widely known, alleviate much confusion. That history is remarkably short, a fact that itself may seem paradoxical.

Modern texts on 'regression analysis' or 'applied linear regression' or 'multiple regression analysis' are almost entirely occupied with examining the use of the method of least squares to fit linear relationships to multivariate data, often for predictive purposes. These texts are based on a statistical methodology that dates back to at least 1805 and the work of Legendre and Gauss and Laplace, methods that were in part foreshadowed by developments a half-century before that.[2] Yet the name 'regression' itself and the concept I discuss here only date from the period 1877–85, and those same texts on 'regression analysis' discuss that concept only sparsely, if at all.

2 The concept of regression

Regression can be viewed as a purely mathematical phenomenon or as an intrinsically statistical concept; to begin with, let us consider how it can be expressed verbally, mathematically, and geometrically, since all of these can be traced to the early days of the concept.

Address for correspondence: SM Stigler, Department of Statistics, University of Chicago, Chicago, IL 60637, USA.
E-mail: stigler@galton.uchicago.edu

0962-2802(97)SM129RA

Verbally, we may consider a stochastic time-varying phenomenon, where two correlated measurements are taken of the same person or object at two different times. For example, we might consider the scores recorded on two examinations taken by the same individual at two separated times. Suppose the first score is exceptionally high – near the top of the class. How well do we expect the individual to do on the second test? The answer, regression teaches us, is 'less well', relative to the class's performance. And the reasoning is clear: there is a selection effect. The high score on the first occasion is surely due to some mixture of successes in two components, to a high degree of skill (a permanent component) and to a high degree of luck (a transient component). The relative bearings of the two components of skill and luck on the first-time score would require measurement to pin down, but the fact that we expect both to have, on average, contributed to the exceptional first outcome is intuitively plausible, even obvious. And on the second occasion we expect the permanent component of skill to persist (for that is the meaning of permanent) while the transient component of luck will, on average, not be present (for that is the meaning of transient). We would not expect that the 'luck' on the second occasion will be bad luck; it may even be good luck – possibly on rare occasions even better that the first time. But it cannot be counted on to persist, and on average there will be no luck at all, neither good nor bad. And so we will have gone from 'high skill plus good luck' to 'high skill alone', a net decrease; still better than average, but less so than before. We expect (with of course no guarantee) regression towards the average. If the first score were exceptionally low, the situation would be reversed, with regression towards the average from below.

Geometrically, the phenomenon can be seen in terms of one simple picture. Figure 1 shows a bivariate normal density; both variables are standardized and the correlation is 0.5. The solid object pictured, if complete, would have a total volume of 1.0 contained in the space between the surface and the X–Y plane. It has, however, been sliced apart. First, a cross-sectional slice is taken perpendicular to the X–Y plane and

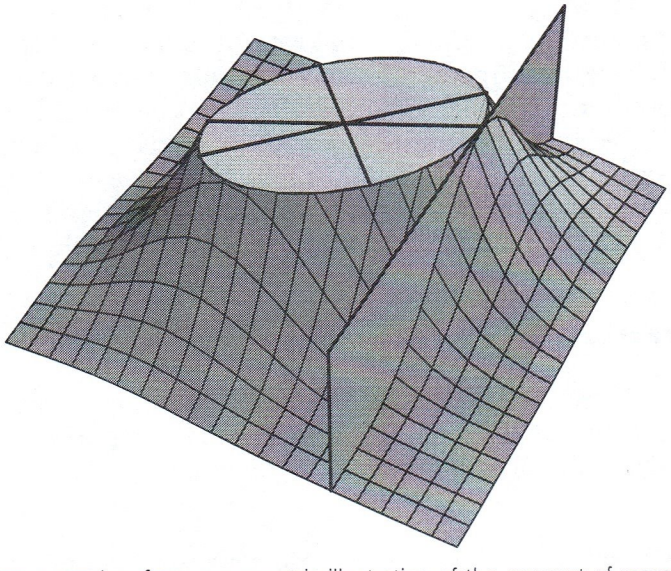


Figure 1 The bivariate normal surface: a geometric illustration of the concept of regression

parallel to the Y axis, intersecting the X-axis at $X = x > 0$, which might be taken as the exceptionally high first-occasion score. Next, the surface is decapitated parallel to the X–Y plane, such that the level curve of intersection (an ellipse) is exactly tangent to the curve of intersection of the first slice (which is a curve proportional to a normal density, the conditional density of Y given $X = x$). The major and minor axes of the ellipse are shown (they are the lines $Y = X$ and $Y = -X$), as is the line from the origin through the point of tangency of the two curves. This latter line is the line of the conditional expectation of Y given $X = x$ (this is clear since it must pass through the mode of the conditional density of Y given X, and for the symmetrical normal distributions, the mode, the median. and the mean must all agree). Then in terms of this diagram the regression phenomenon consists of the obvious observation that the line of conditional expectations must be closer to the X-axis than is the major axis of the ellipse – for it would be clearly impossible for the first slice to touch the ellipse at the point the major axis crosses it, unless the ellipse were collapsed to a line segment, as would only be true if the correlation were 1.0. And so, unless there is perfect correlation between X and Y there must be regression towards the average.

Mathematically, there are several different, equivalent ways of deriving the regression phenomenon.

1) You may begin with two standard normal random variables X and Y with correlation ρ and bivariate density

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp - \left(\frac{1}{2(1-\rho^2)} (x^2 - 2\rho xy + y^2) \right)$$

Then after some algebra the conditional density of Y given $X = x$ is found to be

$$f(y|x) = \frac{f(x,y)}{f_x(x)}$$

$$= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2} \left(\frac{y - \rho x}{\sqrt{1-\rho^2}} \right)^2 \right)$$

which we recognize as the density of a $N(\rho x, 1-\rho^2)$ random variable. Hence the conditional expectation of Y given $X = x$ is ρx, representing regression from x towards the mean of 0.

2) The verbal description given earlier can be expressed mathematically. We may represent

$$X = S + E_1$$
$$Y = S + E_2$$

where S, E_1, and E_2 are independent, S is the 'persistent' trait and the E_i are the 'transient' traits. For the simplest form of the argument, suppose that S and the E_i all have the same distribution, with $E(S) = 0$ and $E(E_i) = 0$. Then

$$\begin{aligned} E(X|Y=y) &= E(S+E_1|S+E_2=y) \\ &= E(S|S+E_2=y)+E(E_1|S+E_2=y) \\ &= E(S|S+E_2=y)+E(E_1) \quad \text{(by independence)} \\ &= E(S|S+E_2=y) \end{aligned}$$

But

$$\begin{aligned} y = E(S+E_2|S+E_2=y) &= E(S|S+E_2=y) \\ &+ E(E_2|S+E_2=y) = 2E(S|S+E_2=y) \end{aligned}$$

and so $E(X|Y=y) = 0.5y$. Note that this argument does not require normality or even the existence of second moments, although if the correlation exists we would clearly have $\rho = 0.5$, in agreement with (1).

3) A different approach is not in terms of standardized variables, but rather is framed sequentially, in terms of a conditional distribution. Let X have a normal distribution $N(0, c^2)$, and let $Y = X + Z$, where Z is $N(0, b^2)$, independent of X. Then Y is $N(0, b^2 + c^2)$ and the correlation of X and Y is

$$\rho = \rho_{XY} = \frac{c^2}{\sqrt{c^2(b^2+c^2)}} = \frac{c}{\sqrt{b^2+c^2}}$$

Clearly the conditional expectation of Y given $X = x$ is simply x; what is the conditional expectation of X given $Y = y$? Finding the bivariate distribution of X and Y and employing a derivation such as that in (1) above tells us that $E(X|Y=y)$ is not y, but rather it is $[(c^2)/(b^2+c^2)]y$, clearly closer to the mean of 0 than is y. The fact that $E(Y|X=x)$ *is* equal to x (rather than being itself closer to the mean of 0) is a reminder that 'regression towards the mean' need literally be true only when the variables are standardized to have the same variances. If we rescale Y to have the same variance as X, by $Y' = \rho Y$, then

$$E(Y'|X=x) = \rho x \text{ and } E(X|Y'=y) = \rho y$$

3 Galton and regression

Francis Galton discovered the phenomenon of regression. Few conceptual advances in statistics can be as unequivocally associated with a single individual. Least squares, the central limit theorem, the chi-squared test – all of these were realized as the culmination of many years of exploration by many people. Regression too came as the culmination of many years' work, but in this case it was the repeated efforts of one individual.

The first glimmers of the idea can be found already in Galton's 1869 book *Hereditary genius*. In that work he studied the way talent ran in families, and most of the book consists of lists of eminent people and their eminent relatives – great scientists and their kin with known scientific accomplishments (e.g. the Bernoullis), musicians and

their musical kin (e.g. the Bachs), and so forth. But despite the inevitable arbitrariness in his classifications and evaluation of eminence, Galton noted that there was a marked tendency for a steady decrease in eminence the further down or up the family tree one went from the great man (e.g. Jacob Bernoulli or Johann Sebastian Bach) whose fame led to the family's inclusion in the study. Even with dogs this was true: 'If a man breeds from strong, well-shaped dogs, but of mixed pedigree, the puppies will be sometimes, but rarely, the equals of their parents. They will commonly be of a mongrel, nondescript type, because ancestral peculiarities are apt to crop out in the offspring.'[3]

In 1869 Galton only vaguely approached the concept in its verbal form, but he was unable to formulate in a precise way how the accidental 'cropping out' of 'ancestral peculiarities' might be encompassed in a theory. Still the question kept gnawing at him; over the years 1874–88 he revisited this problem repeatedly, and, bit by bit, he overcame it in one of the grand triumphs of the history of science. The story is an exciting one, involving science, experiment, mathematics, simulation, and one of the great mental experiments of all time. But it is a long story, one I have examined in detail in my book,[2] and so I shall only relate it in outline here.

In the years 1874–77, Galton launched his first assault upon this conundrum: how and why was it that talent or quality once it occurred tended to dissipate rather than grow. He never lost interest in the study of the inheritance of human genius, but he realized early on that intellectual quality was not an area that permitted either easy measurement on a wide scale or active experimentation. And so he fell back on studies of other measurable qualities, particularly stature – height – in humans, and he began a series of experiments involving the measurement in successive generations of the diameter of sweet peas. And while considering these experiments, he invented a wonderful machine, the Quincunx, that was to serve as an analogue for hereditary processes and provide the key insight to the solution.

Galton had been puzzled by how to reconcile the standard theory of errors with what he observed and knew to be true from experiments. The theory of errors held that a normal population distribution would be produced through the accumulation of a large number of small accidental deviations, and there seemed to be no other way to account for the ubiquitous appearance of that normal outline. Galton's experiments with sweet peas and his studies of human stature agreed with earlier work by the Belgian statistician Adolphe Quetelet: the world, by and large, was normally distributed. Yet, as Galton realized, this did not square with the fact that in heredity there were large and important causes of deviations at work: inheritance of talent. height, or diameter was not perfect, but these qualities *did* run in families. The normal distribution he and others found was not the exclusive result of small accidental causes; it had somehow to be reconciled with the influence of the large and invariable causes of heredity.

In 1873 Galton had a tradesman make for him a machine he called the Quincunx. It consisted of a board with a funnel at the top through which lead shot could be released to fall through a succession of offset rows of pins, collecting at the bottom in vertical compartments (for a photograph of the original machine see Stigler,[2] p 277). The left panel of Figure 2 shows a schematic rendition. The name 'Quincunx' was derived from the similarity of the pattern of pins to the arrangement of cultivated fruit trees in

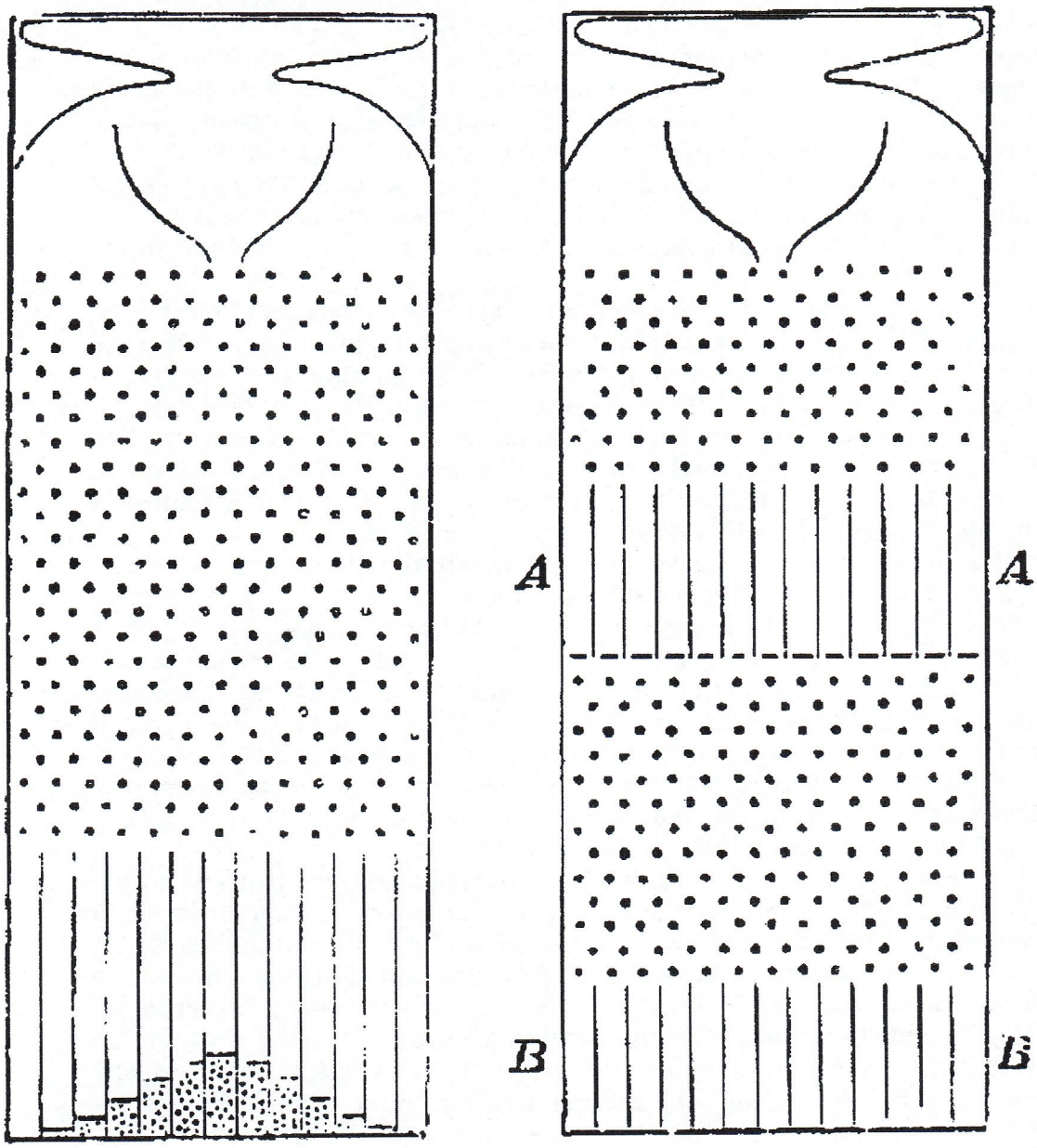


Figure 2 A schematic drawing of Galton's Quincunx[4]

English agriculture, a pattern that was known as quincunxial because it was based on a square of four trees with a fifth in the centre. Galton's Quincunx was initially intended to illustrate the workings of a large number of small accidental causes to produce a normal-like distribution. It might be likened to a dynamic version of Pascal's triangle: As shot pass from top to bottom they are randomly deflected at each row, and if the machine is well made and in balance the shot will produce an outline at the bottom

where the number of shot in each compartment is proportional to the number of paths to that compartment. That is, the number of shot in a compartment will be proportional to the binomial coefficients – a nearly normal distribution if the number of rows of pins is at all large.

The Quincunx illustrated the manner in which a large number of small accidents could produce a normal distribution. But what of the large and not-so-accidental causes that Galton found inherited to one degree or another in his studies? The evolutionary progress of the shot through the Quincunx led Galton to his fundamental first insight through one of the great mental experiments in the history of science. I term this a mental experiment because, while Galton clearly in several places described the variant of the Quincunx that performed the experiment, there is no indication that he actually built the apparatus. And having tried to build such a machine, I can testify that it is exceedingly difficult to make one that will accomplish the task in a satisfactory manner. Galton first imagined taking the Quincunx apart in the middle and stretching it out, but to ensure that the stretching does not alter the final distribution of the shot he would add vertical barriers to keep them from straying while they traversed the gap. Galton's printed diagram from 1889 is shown in the second panel of Figure 1; he illustrated the idea in correspondence as early as 1877 (see Stigler,[2] pp 278–79).

Clearly with these barriers the introduction of the gap would have no effect on the distribution of shot among the compartments at the bottom. Galton then conceived of introducing a barrier at the bottom of the gap, turning the barriers into a second set of compartments like those at the bottom. What effect would that have? Again it is clear that all this would do is to foreshorten the Quincunx; with fewer rows of pins to traverse, the shot would still come to rest in a normal-like distribution, but one that was less disperse than if they had been allowed to finish the course. Galton would then release the shot from this midlevel, but only from one compartment: this would be expected to produce a small normal distribution immediately below the compartment from which they were released. Proceed then to release the remaining compartments, one at a time. Each will produce its own little normal curve; those near the centre being larger than those more extreme, because more shot will have been deposited in the central compartments by the first stage of the Quincunx. And when all have been released, the result – the sum of all the little normal curves – will be as if no interruption at all had taken place!

Galton's imagination had shown how the normal world could be dissected into components, components which could be traced back to the location of the shot at the end of a first stage. The machine was a beautiful match to his investigations of inheritance. The seeming homogeneity of the final outline could be seen now as a mixture derived from previous generations. Indeed, Galton's mental experiment can be interpreted as an analogue proof of the mathematical theorem, that a normal mixture of normal distributions is itself normal (or, in the discrete version, that a convolution of binomial distributions with the same p is binomial). You can even see the phenomenon of regression: the expected final position of a shot released from the mid-level is immediately below it, but what is the expected origin of a shot on the bottom level? Clearly towards the centre from its position, since there are more shot originating towards the centre than further away.

Table 1 One of Galton' s correlation tables (from Francis Galton, Family likeness in stature, *Proceedings of the Royal Society of London* 1886; **40**: 42–73). Galton's 1885 crosstabulation of 928 'adult children' born of 205 mid-parents, by their height and their mid-parent's height

Heights of the Mid-parents in inches.	Heights of the Adult Children.														Total Number of		Medians.
	Below	62·2	63·2	64·2	65·2	66·2	67·2	68·2	69·2	70·2	71·2	72·2	73·2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72·5	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72·2
71·5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69·9
70·5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69·5
69·5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68·9
68·5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68·2
67·5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67·6
66·5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67·2
65·5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66·7
64·5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65·8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66·3	67·8	67·9	67·7	67·9	68·3	68·5	69·0	69·0	70·0	..	..	..	..	..

With the Quincunx in mind. Galton's later correlation tables take on a whole new meaning. For example, in Table 1 the right-hand column 'Total no. of adult children' is seen as the distribution of the shot at the mid-level, the rows of counts as the corresponding little normal curves, and the 'Totals' of the bottom row as the final outline of the Quincunx.

Even by 1877 Galton had begun to assemble these insights mathematically. He had empirically noted the tendency for 'reversion' towards the mean and labelled this 'r'. In his notation, if c = the dispersion (essentially, standard deviation) of the first generation, d = the dispersion of the second generation and v = the dispersion of the offspring (the little normal curves), then since the position of a second generation individual was the sum of its 'reverted' average displacement from its parent (say rz, where z was the first generation position) and its random deviation from that position, these dispersions would be related by $d^2 = v^2 + r^2c^2$. But why did the reversion take the linear form rz? And why would the population dispersion remain stable; that is, what mechanism produced $d = c$?

The answer to this (that $d = c$ was a necessary consequence of population stability) did not come to Galton until 1885, when, inspired by tables such as Table 1, and with a slight assist from the Cambridge mathematician JH Dickson, he produced a full formulation in terms of the bivariate normal distribution. He summarized and elaborated upon this formulation in his 1889 *Natural inheritance*.[4] His discussion there included the geometric interpretation of regression and the mathematical formulation given earlier as (3) (which we can recognize now as a description of the working of the Quincunx, with X = the reverted first generation position and Z is the displacement of offspring from parent), and much more. He was aware that there were two regression lines. He even described a variance components model for fraternal relationships, and he discussed how to estimate the components of variance. By the time *Natural inheritance* appeared, he had, while considering problems in physical anthropology and

forensic science, noticed that when two variables were expressed in standardized units, the two regression lines had the same slope, and he suggested using that slope, which he termed the 'index of co-relation,' as a measure of the strength of the relationship. He interpreted the correlation coefficient both as a regression coefficient and as what we would now term an intraclass correlation coefficient.[5]

4 The understanding of the regression phenomenon

It is fair to say that by 1889 Francis Galton had a clear understanding of the concept of regression. He did not have the command of all the mathematical apparatus I used in the discussion of the concept early in this essay, but his written discussion captured the essence of all of the different formulations given, and his mathematics reflected at least that of (2) and (3). Regression was no longer simply an empirical observation, it was a mathematical deduction. He wrote (p 95)

> However paradoxical it may appear at first sight, it is theoretically a necessary fact, and one that is clearly confirmed by observation, that the Stature of the adult offspring must on the whole be more *mediocre* than the stature of their Parents.[4]

Questions about how to best estimate the coefficients of the problem, the correlation coefficient and the parameters of the bivariate normal distribution, would be addressed later by Francis Edgeworth and Karl Pearson, but Galton's grasp of the concepts was as firm as any you are likely to encounter even today. Galton himself was naive in assuming that if data were recorded on a sequence of occasions (not only two) that regression necessarily continued, even at the same rate. Karl Pearson named this 'Galton's Law of Ancestral Heredity,' and even Pearson did not seem to appreciate that the continuation of the phenomenon after the first generation requires rather special assumptions.[6] How well did Galton do in communicating that understanding? If judged by the way he is received by a reader a century after he wrote, the answer would have to be, very well indeed. He wrote in clear and direct prose, in terms that we can understand, with the penetration and clarity that are characteristic of only some of the greatest minds. But that is not the standard that is called for. How well did his contemporaries understand his message?

Statisticians generally grasped the concept quite well at one level. Edgeworth and Pearson set to work developing the mathematics of regression further, moving towards multiple dimensions and exploring optimum procedures for estimating correlation. In 1901 Bowley wrote the earliest English text to include the new statistical methods, and he included a chapter on the mathematics of the bivariate normal distribution, including both lines of conditional expectation.[7] A reader who came away from Bowley's discussion with the impression that the primary importance of regression was for the study of evolution should have been excused, however, Udny Yule incorporated a full appreciation of the idea of two regression lines into his highly influential text from the first edition.[8] At least one perceptive early reviewer of Galton, the philosopher John Dewey, called specific attention to the phenomenon of regression, even noting in effect its dependence upon a stationary population, when he wrote that it might not hold in the inheritance of wealth: 'The tendency of wealth to breed wealth, as illustrated by any interest table, and the tendency of extreme poverty to induce

conditions which plunge children still deeper into poverty, would probably prevent the operation of the law of regression toward mediocrity.'[9] Of course Galton could have replied that even then the law would hold in standardized units.

Still, there were clearly limitations to the general understanding of regression as a phenomenon capable of dangerously misleading. The biometrician Frank Weldon, who himself had a very good grasp of Galton's message, wrote in a 1905 lecture that

> [T]his phenomenon of regression ... is not generally understood [V]ery few of those biologists who have tried to use [Galton's] methods have taken the trouble to understand the process by which he was led to adopt them, and we constantly find regression spoken of as a peculiar property of living things, by virtue of which variations are diminished in intensity during their transmission from parent to child, and the species is kept true to type. This view may seem plausible to those who simply consider that the mean deviation of children is less than that of their fathers: but if such persons would remember the equally obvious fact that there is also a regression of fathers on children, so that the fathers of abnormal children are on the whole less abnormal than their children, they would either have to attribute this feature of regression to a vital property by which children are able to reduce the abnormality of their parents, or else to recognize the real nature of the phenomenon they are trying to discuss.[10]

In the decades after Weldon wrote, the situation did not change. Following Yule, regression was a staple of textbooks. Its mathematics could be said to be well understood by mathematical statisticians, while applied statisticians, if they were aware of it at all, thought of it as either the use of the method of least squares or as only a biological process. The term 'regression' soon came to be regarded as archaic, often accompanied by a brief explanation of its roots in biology but with no indication of the relevance of those roots to other applications. In 1924 the economic statistician Frederick C Mills could write, 'The term is now used generally, as indicated above, though the original meaning has no significance in most of its applications.'[11]

It was therefore a trap waiting for the unwary, who were legion. The most spectacular instance of a statistician falling into the trap was in 1933, when a Northwestern University professor named Horace Secrist unwittingly wrote a whole book on the subject, *The triumph of mediocrity in business*.[12] In over 200 charts and tables, Secrist 'demonstrated' what he took to be an important economic phenomenon, one that likely lay at the root of the great depression: a tendency for firms to grow more mediocre over time. Secrist was aware of Galton's work; he cited it and used Galton's terminology. The preface even acknowledged 'helpful criticism' from such statistical luminaries as HC Carver (the editor of the *Annals of Mathematical Statistics*), Raymond Pearl, EB Wilson, AL Bowley, John Wishart and Udny Yule. How thoroughly these statisticians were informed of Secrist's work is unclear, but there is no evidence that they were successful in alerting him to the magnitude of his folly (or even if they noticed it). Most of the reviews of the book applauded it.[13–15] But there was one dramatic exception: in late 1933 Harold Hotelling wrote a devastating review, noting among other things that 'The seeming convergence is a statistical fallacy, resulting from the method of grouping. These diagrams really prove nothing more than that the ratios in question have a tendency to wander about.'[16] Secrist did not understand the criticism, leading Hotelling to reiterate the lesson in a subsequent letter in even plainer language: 'When in different parts of a book there are passages from which

the casual reader may obtain two different ideas of what the book is proving, and when one version of the thesis is interesting but false and the other is true but trivial, it becomes the duty of the reviewer to give warning at least against the false version.'[17]

One would think that so public a flogging as Secrist received for his blunder would wake up a generation of social scientists to the dangers implicit in this phenomenon, but that did not happen. Textbooks did not change their treatment of the topic, and if there was any increased awareness of it, the signs are hard to find. In the more than two decades between the Secrist–Hotelling exchange in 1933 and the publication in 1956 of a perceptively clear exposition in a textbook by W Allen Wallis and Harry Roberts, I have only encountered the briefest acknowledgements.[18] A paper in *Psychometrika* by RL Thorndike[19] is an exception; like Hotelling's review, Thorndike's paper was a reaction to blunders in the literature. Thorndike disclaimed originality ('It is not the purpose of this paper to present any scintillating new statistical ideas'), but he clearly expected that his tutorial would be news to many readers. He mentioned only one offender, a psychologist from the University of Iowa named Crissey, but stated, 'I select this example without malice – I might have selected any of a number of others'. The more common rule over this two-decade period was that textbooks such as the successive revisions of Yule's book by MG Kendall kept repeating the earlier material with more recent references and enhanced mathematics.

Even after 1956, when (perhaps influenced by Wallis and Roberts) the topic attracted increasing attention, blunders persisted. In 1970 a political economist AO Hirschman (who had presumably not read Hotelling's review, and was evidently innocent of any awareness of the regression phenomenon) cited Secrist's book, writing, 'An early, completely forgotten empirical work with a related theme has the significant title *The triumph of mediocrity in business*, by Horace Secrist, ... The book contains an elaborate statistical demonstration that, over a period of time, initially high-performing firms will on the average show deterioration while the initial low performers will exhibit improvement.'[20] Some writers have known of the problem but still fallen in the trap (see, for example, Friedman[21] for a discussion of two of these). Other researchers who have known of the phenomenon but not understood it have been frightened by the spectre of one type of error into making another: in at least one instance,[22] researchers were so worried about the possibility of committing the fallacy that they introduced a correction for 'regression effects' where, not only was none needed, the 'correction' produced an erroneous result!

The recurrence of regression fallacies is testimony to its subtlety, deceptive simplicity, and, I speculate, to the wide use of the word regression to describe least squares fitting of curves, lines, and surfaces. Researchers may err because they believe they know about regression, yet in truth have never fully appreciated how Galton's concept works. History suggests that this will not change soon. Galton's achievement remains one of the most attractive triumphs in the history of statistics, but it is one that each generation must learn to appreciate anew, one that seemingly never loses its power to surprise.

Acknowledgements

I am grateful to William Kruskal, IJ Good and Xuming He for references and comments, and to Miller Maley for preparing Figure 1. This work was supported in part by the NSF.

References

1 Friedman M. Do old fallacies ever die? *Journal of Economic Literature* 1992; **30**: 2129–32.
2 Stigler SM. *The history of statistics*. Cambridge, MA: Harvard University Press, 1986.
3 Galton F. *Hereditary genius*. London: Macmillan 1869: 64.
4 Galton F. *Natural inheritance*. London: Macmillan, 1889.
5 Stigler SM. Francis Galton's account of the invention of correlation. *Statistical Science* 1989; **4**: 73–86.
6 Nesselroade J, Stigler SM, Baltes P. Regression toward the mean and the study of change. *Psychological Bulletin* 1980; **87**: 622–37.
7 Bowley AL. *Elements of statistics*. London: PS King, 1901: 316–26, and later editions.
8 Yule GU. *An introduction to the theory of statistics*. London: Charles Griffin, 1911 (and many later editions).
9 Dewey J. Galton's statistical methods. *Publications of the American Statistical Association* 1889; **7**: 331–34. [Quoted in Stigler, *The history of statistics*. 1986: 301, and at more length in Stigler, A look backward on the occasion of the centenary of JASA, *Journal of the American Statistical Association* 1988; **83**: 583–87.]
10 Strong TB ed. *Lectures on the method of science*. Oxford: Clarendon Press, 1906: 106–107.
11 Mills FC. *Statistical methods. Applied to economics and business*. New York: Henry Holt, 1924: 394.
12 Secrist H. *The triumph of mediocrity in business*. Evanston, IL: Bureau of Business Research, Northwestern University. 1933.
13 Elder RF. Review of *The triumph of mediocrity in business* by Secrist H. *American Economic Review* 1934: **24**: 121–22.
14 King WI. Review of *The triumph of mediocrity in business* by Secrist H. *Journal of Political Economy* 1934; **42**: 398–400.
15 Riegel R. Review of *The triumph of mediocrity in business* by Secrist H. *Annals of the American Academy of Political and Social Science* 1933; **170**: 178–79.
16 Hotelling H. Review of *The triumph of mediocrity in business* by Secrist H. *Journal of the American Statistical Association* 1933; **28**: 463–65.
17 Secrist H, Hotelling H, Rorty MC. Open letters I. *Journal of the Anierican Statistical Association* 1934; **29**: 196–200; see Stigler SM., The history of statistics in 1933. *Statistical Science* 1996, **11**: 244–52, for a full account of Secrist and Hotelling.
18 Wallis WA, and Roberts H. *Statistics: a new approach*. Glencoe IL: Free Press, 1956: 258–63.
19 Thorndike RL. Regression fallacies in the matched groups experiment. *Psychometrika* 1942; **7**: 85–102.
20 Hirschman AO. *Exit, voice, and loyalty: responses to decline in firms, organizations, and states*. Cambridge, MA: Harvard University Press, 1970.
21 Friedman M. Do old fallacies ever die? *Journal of Economic Literature* 1992; **30**: 2129–32.
22 Stigler SM. Psychological functions and regression effect. *Science* 1979; **206**: 1430.

From The Collected Works of Milton Friedman, compiled and edited by Robert Leeson and Charles G. Palm.

"Do Old Fallacies Ever Die?"

by Milton Friedman

Journal of Economic Literature 30, December 1992, pp. 2129-2132

In 1933, Harold Hotelling reviewed a book by Horace Secrist, entitled *The Triumph of Mediocrity in Business,* in which Secrist presented evidence purporting to show that business enterprises were tending to converge in size. Hotelling pointed out that Secrist's evidence did not justify his conclusions. For a number of different variables, Secrist had plotted

> averages of groups, arrayed according to the value of the variable in the first year of the series. If the concerns were arrayed according to the values taken by the variable in the last year of the series, the lines would diverge. … The seeming convergence is a statistical fallacy, resulting from the method of grouping. …
>
> The real test of a tendency to convergence would be in showing a consistent diminution of variance, not among means of groups, but among individual enterprises. (Hotelling 1933, p. 464)

In 1991, Jeffrey G. Williamson reviewed a book by William J. Baumol, Sue Anne Batey Blackman, and Edward N. Wolff, *Productivity and American Leadership.* One thesis of the book is that the rates of growth of various countries have tended to converge. As it happens, their thesis appears to be correct, whereas, according to Hotelling, Secrist's was not. But both the reviewer and the book cite evidence for convergence that is tainted by the statistical fallacy that Hotelling called attention to almost sixty years earlier.

To quote from Williamson's review, "Figure 4 [reproduced here as Figure 1] confirms the convergence thesis: The bigger the productivity gap in 1950, the faster the growth 1950–79" (1991, pp. 57–58). The figure does no such thing. Suppose we do as Hotelling suggests and use the terminal year rather than the initial year in plotting the rates of growth. The result is Figure 2. Japan and the U.S. clearly show convergence even on this biased basis. But the other countries show essentially no correlation: the calculated regression coefficient is positive, but not statistically significant. Figure 2 does support the convergence thesis, but only because it shows so much less close a positive correlation than it would if the regression fallacy stressed by Hotelling were alone operative.

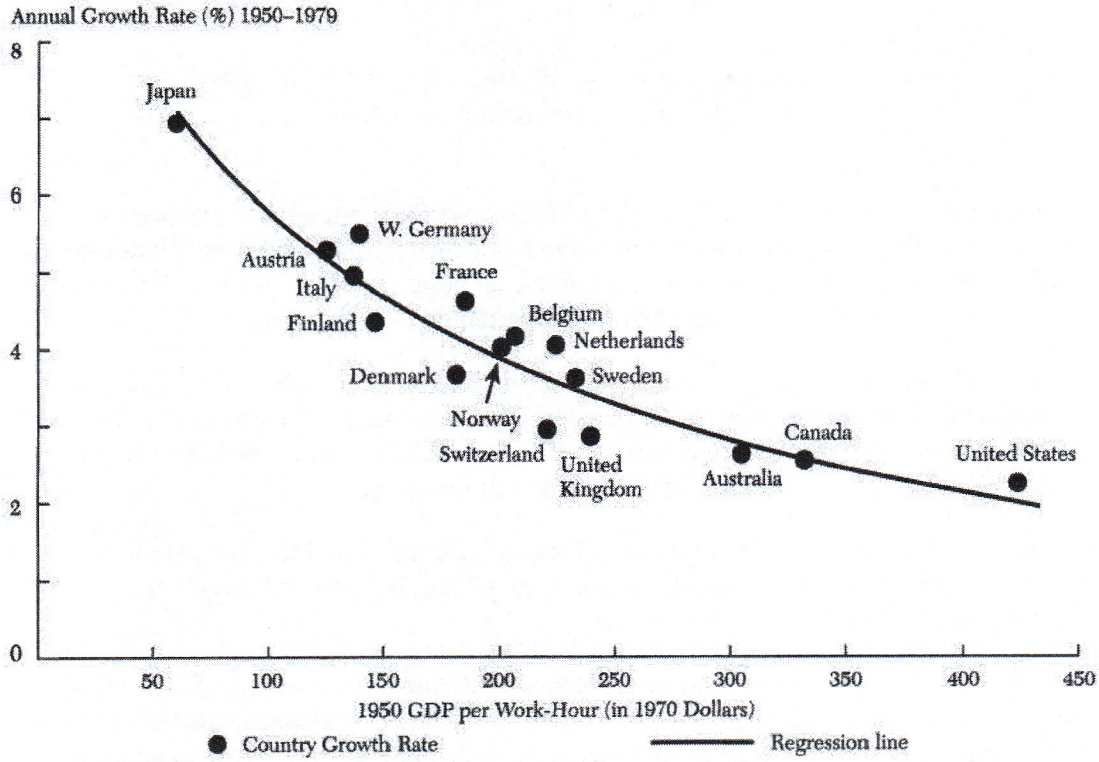


Figure 1. Postwar Productivity (GDP per work hour) Growth Rate, 16 Industrialized Countries

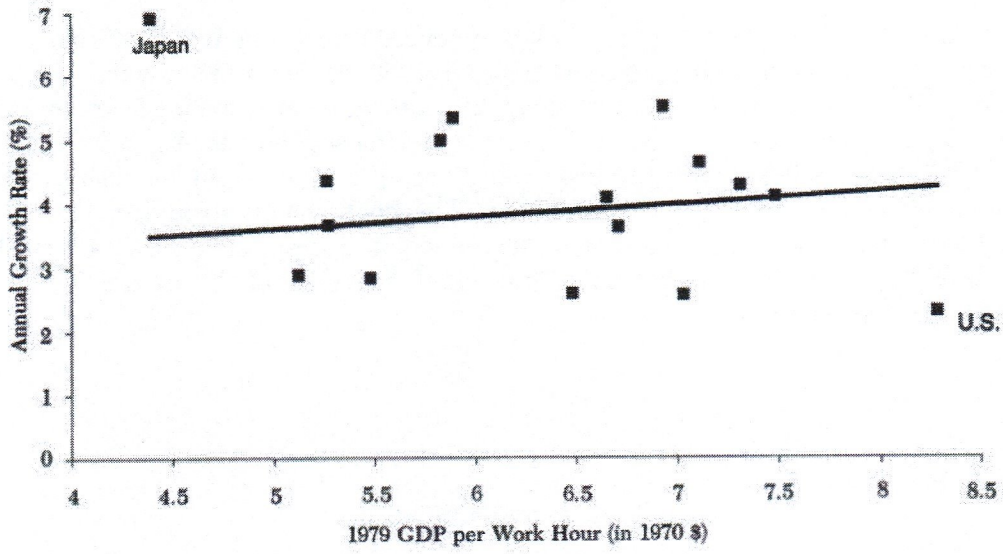


Figure 2. Growth Rate of Productivity vs Terminal Level, 1950–1979

(NOTE: Regression Excludes Japan and U.S.)

The book does, while the review does not, present in Figure 5.2 (Baumol, Blackman, and Wolff 1989), reproduced here as Figure 3, what Hotelling called "The real test of a tendency to convergence." These plots of the coefficient of variation of GDP per work-hour and GDP per capita do show, in Hotelling's words, "a consistent diminution of variance ... among individual" countries. Figures 5.1 and 5.4 in the book are also relevant evidence, untainted by the regression fallacy.[1] But that is not true of three other figures (5.3, 5.5, and 5.6), all of which are tainted.

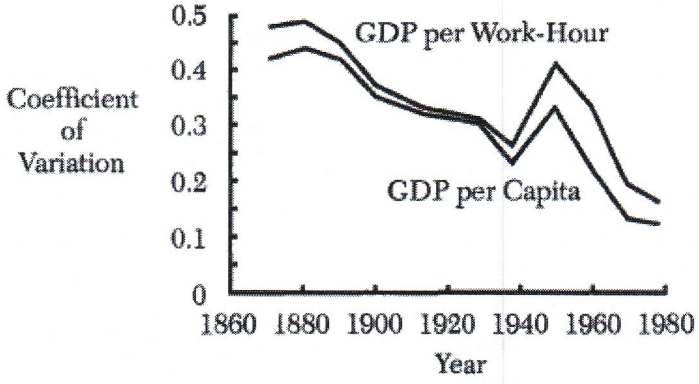


Figure 3. Coefficients of Variation, GDP per Work-Hour and GDP per Capita, 16 Countries, 1870–1979

I find it surprising that the reviewer and the authors, all of whom are distinguished economists, thoroughly conversant with modern statistical methods, should have failed to recognize that they were guilty of the regression fallacy. After all, the phenomenon in question is what gave regression analysis its name.[2] However, surprise may not be justified in light of the ubiquity of the fallacy both in popular discussion and in academic studies.

For example, "everyone knows" that job creation comes mainly from small firms. That proposition may be true but the evidence offered for it that I have seen classifies firms by size in an initial year and traces subsequent levels of employment—precisely what Secrist did. I have yet to see what the data show if firms are classified by their terminal size, or by their average size over a period.

Similarly, in academic studies, the common practice is to regress a variable Y on a vector of variables X and then accept the regression coefficients as supposedly unbiased estimates of structural parameters, without recognizing that all variables are only proxies for the variables of real interest, if only because of measurement error, though generally also because of transitory factors that are peripheral to the subject under consideration. I suspect that the regression fallacy is the most common fallacy in the statistical analysis of economic data, alleviated only occasionally by consideration of the bias introduced when "all variables are subject to error."

As a student of Hotelling not long after his review had been published, I early became aware of the regression fallacy—or, perhaps better, trap. That knowledge came in handy when Simon Kuznets and I studied the relation between incomes of the same individuals in different years. To avoid the trap, I introduced the concept of permanent and transitory components of income, in the process referring to Hotelling's review (Friedman and Kuznets 1945, p. 331).

Later, in my *Theory of the Consumption Function* (1957), the regression fallacy came to my aid in resolving the apparent conflict between budget studies and Kuznets' time series data. It was generally believed that budget studies had demonstrated that the marginal propensity to consume was greater than the average propensity, i.e., that the higher the income, the lower the fraction that would be spent on consumption. Yet Kuznets' time series data showed no such tendency.

The explanation was that budget studies almost invariably classified consumer units by income and then calculated averaged consumption for various income classes. Few if any consumer units were in the lowest income class despite having an unusually high income; many were there because of an unusually low income; their permanent income exceeded their measured income, and conversely at the other end of the scale. Consequently, it was not surprising that units with low measured incomes would spend on consumption a higher fraction of their measured income than units with high measured income. The only budget study for which I could get average income for units classified by consumption spending showed the expected reverse relation. Indeed, the regression fallacy was the seed out of which my permanent income hypothesis grew (Friedman 1957, ch. 3, and pp. 201–02).[3]

Williamson's review and the book he reviewed are excellent vehicles for demonstrating the difficulty of rooting out appealing fallacies, precisely because both are of unusually high quality, and are marred in only a minor way by having fallen into the regression trap.

References

Baumol, William J.; Blackman, Sue Anne Batey And Wolff, Edward N. *Productivity and American leadership: The long view.* Cambridge and London: MIT Press, 1989.

Friedman, Milton. *A theory of the consumption function.* Princeton NJ: Princeton U. Press, 1957.

Friedman, Milton and Kuznets, Simon. *Income from independent professional practice.* NY: National Bureau of Economic Research, 1945.

Friedman, Milton and Schwartz, Anna J. *Monetary trends in the United States and the United Kingdom.* Chicago: U. of Chicago Press, 1982.

Hotelling, Harold. "Review of *The triumph of mediocrity in business,* by Horace Secrist," *J. Amer. Statist. Assoc.,* Dec. 1933, 28(184), pp. 463–65.

Williamson, Jeffrey G. "Productivity and American Leadership: A Review Article," *J. Econ. Lit.,* Mar. 1991, 29(1), pp. 51–68.

Notes

[1] Moses Abramovitz has called my attention to a significant qualification of Hotelling's "real test": it implicitly assumes that measurement error (or transitory variance) is the same in the years compared. That qualification may be minor for the data in Figures 2 and 3 for the post-World War II period. However, improvements in the accuracy of the data may well account for a significant part, even if not all, of the decline in dispersion in Figure 3 from 1860–1940.

[2] Galton examined the heights of fathers and sons, and found that the sons of tall fathers tended to be shorter than their fathers, i.e., regressed toward the mean; similarly, the fathers of tall sons tended to be shorter than their sons, i.e., regressed toward the mean. This is simply the well-known phenomenon that in a linear regression of x and y, the regression of y on x is flatter relative to the x axis than the regression of x on y.

[3] Still later, in our book *Monetary Trends,* Anna Schwartz and I allowed for the regression effect by systematically reporting upper and lower limits of computed parameters derived by reversing independent and dependent variables in regressions. See Friedman and Schwartz (1982, esp. ch. 5, fns. 28 and 29, pp. 173–74; ch. 6, p. 227).

10/1/12

Lecture of 26.03.20

giovedì 26 marzo 2020 11:13

$\underline{X}$ r.v. $\sim N_p(\mu, \Sigma)$

Prop If $\det(\Sigma) > 0$

$$(\underline{X}-\mu)^T \Sigma^{-1} (\underline{X}-\mu) \sim \chi^2(p).$$

proof

Remember: if $Z_1, \dots, Z_p$ iid $\sim N_1(0,1)$

$\sum_{i=1}^{p} Z_i^2 \sim \chi^2(p)$.

Consider: $\Sigma = \sum \lambda_i e_i e_i^T$ spect. decomp.

$P = [e_1 \cdots e_p] \quad \Lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{bmatrix}$

$\Sigma = P \Lambda P^T$

Consider: $\Lambda^{-1/2} P^T (\underline{X}-\mu)$ → linear transform

$\Lambda^{-1/2} P^T (\underline{X}-\mu) \sim N_p(\underline{0}, \Lambda^{-1/2} P \Sigma P^T \Lambda^{-1/2})$

$\Lambda^{-1/2} P \Sigma P^T \Lambda^{-1/2} =$

$= \Lambda^{-1/2} \underbrace{P^T P}_{\text{orthogonal}} \Lambda \underbrace{P^T P}_{\text{orthogonal}} \Lambda^{-1/2} =$

$= \Lambda^{-1/2} \Lambda \Lambda^{-1/2} = I$

$\Rightarrow \Lambda^{-\frac{1}{2}} P^T (\underline{X}-\mu)$ is gaussian (since it's a linear comb. of gaussian variables)

$\underline{Z} = \Lambda^{-1/2} P^T (\underline{X}-\mu) \sim N_p(\underline{0}, I)$

$\underline{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_p \end{pmatrix} \Rightarrow Z_1 \dots Z_p$ iid $\sim N_1(0,1)$

$(\underline{X}-\mu)^T \Sigma^{-1} (\underline{X}-\mu) =$

$= (\underline{X}-\mu)^T P \Lambda^{-1/2} \Lambda^{-1/2} P^T (\underline{X}-\mu)$

$= \underline{Z}^T \underline{Z} = \sum_{i=1}^{p} Z_i^2 \sim \chi^2(p)$ ■

Obs: If $\det(\Sigma) = 0$ let $k = \text{rank}(\Sigma)$ ← $\det(\Sigma) = 0 \equiv \prod_{i=1}^{p} \lambda_i = 0 \equiv \exists i : \lambda_i = 0$

$\Sigma = \sum_{i=1}^{p} \lambda_i e_i e_i^T \qquad \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k > 0 = \lambda_{k+1} = \dots = \lambda_p$

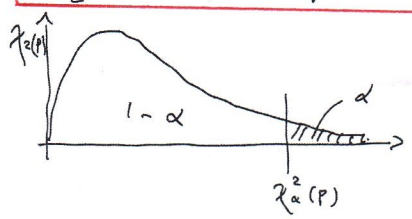
we suppose the last ones are 0

GENERALIZED INVERSE $\Sigma^- = \sum_{i=1}^{k} \frac{1}{\lambda_i} e_i e_i^T$

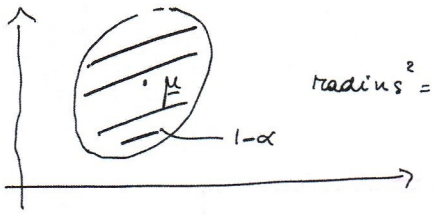
Prove that: $(\underline{X}-\mu)^T \Sigma^- (\underline{X}-\mu) \sim \chi^2(k)$.

Corollary. $\alpha \in (0,1)$, $\det(\Sigma) > 0$

$$P_\Sigma\left[(\underline{X}-\mu)^T \Sigma^{-1} (\underline{X}-\mu) \leq \chi^2_\alpha(p) \right] = 1-\alpha$$



If we take a gaussian distribution centered in $\underline{\mu}$ and we take the points with the **Mahalanobis's** distance from $\underline{\mu} = \sqrt{\chi^2_\alpha(p)}$ then within this region there is a $1-\alpha$ probability.



Up to now we worked with the model, now we have to estimate the parameters

Estimators of $\underline{\mu}$ and Σ for $N_p(\underline{\mu}, \Sigma)$.

$$\mathbb{X} = \begin{bmatrix} \underline{x}_1^T \\ \vdots \\ \underline{x}_n^T \end{bmatrix} \quad \text{data}$$

$\underline{x}_i^T \xleftarrow{\text{obs}} \underline{X}_i$ random vector

$\underline{X}_1, \dots, \underline{X}_n$ iid $\sim N_p(\underline{\mu}, \Sigma)$

$\underline{\mu}, \Sigma$ are unknown.

- $\bar{\underline{X}} = \frac{1}{n} \sum_{i=1}^{n} \underline{X}_i$
- $S = \frac{1}{n-1} \sum_{i=1}^{n} (\underline{X}_i - \bar{\underline{X}})(\underline{X}_i - \bar{\underline{X}})^T$

} unbiased $\underline{\mu}$, Σ

MLE estimators for $\underline{\mu}$ and Σ:

maximum likelihood estimator

$$\text{"}\mathbb{P}\text{"}\left(\underline{X}_1 = \underline{x}_1, \dots, \underline{X}_n = \underline{x}_n\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)^p \, \text{Det}(\Sigma)}} \exp^{-\frac{1}{2}(\underline{x}_i - \underline{\mu})^T \Sigma^{-1} (\underline{x}_i - \underline{\mu})} \left(d\underline{x}_1 \cdots d\underline{x}_n\right)$$

we're not really considering the prob. to be exactly one value (since it's a continuous distribution), we're considering a value in an infinitesimal neighborhood

L likelihood : $(\underline{\mu}, \Sigma) \longrightarrow \text{"}\mathbb{P}_R\text{"}[\underline{X}_1 = \underline{x}_1, \dots, \underline{X}_n = \underline{x}_n]$ ←

given $\underline{X}_1 = \underline{x}_1, \dots, \underline{X}_n = \underline{x}_n$

L likelihood.

! given the data we want a function that maps any possible value of the parameters to the probability of observing that data if those params. were the true ones.

"Likelihood of the parameters given the data we observed"

Teo

$$\underset{(\underline{\mu}, \Sigma):\ \underline{\mu} \in \mathbb{R}^p,\ \Sigma\ p\times p\ \text{pos. def}}{\arg\max} \; L(\underline{\mu}, \Sigma \mid \underline{X}_1 = \underline{x}_1 \dots \underline{X}_n = \underline{x}_n) = (\hat{\underline{\mu}}, \hat{\Sigma})$$

where : $L(\underline{\mu}, \Sigma \mid \underline{X}_1 = \underline{x}_1 \cdots \underline{X}_n = \underline{x}_n)$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)^p \, \text{Det}(\Sigma)}} \exp\left[-\frac{1}{2}(\underline{x}_i - \underline{\mu})^T \Sigma^{-1} (\underline{x}_i - \underline{\mu})\right]$$

$$\Rightarrow \quad \hat{\underline{\mu}} = \bar{\underline{X}}$$

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (\underline{X}_i - \bar{\underline{X}})^T (\underline{X}_i - \bar{\underline{X}}) = \frac{n-1}{n} S$$

(biased). → maybe we want an estimator which is good today, not on average (like the unbiased one) (**NOTICE** that with longe n there is no big difference with S)

Finding the maximum likelihood estimator means finding the values of the parameters so that the likelihood is maximized (≡ Finding the values of the parameters so that what we have observed has the max probability of beign observed)

Properties of MLE's

$\theta \in \mathbb{R}^k$ parameter (ex. $\theta = (\underline{\mu}, \Sigma)$)

$\hat{\theta} = \hat{\theta}(\text{data})$ is MLE estimator of θ

$h : \mathbb{R}^k \longrightarrow \mathbb{R}^j$

? $\widehat{h(\theta)}$ MLE ?

Invariance property of MLE: $\widehat{h(\theta)} = h(\hat{\theta})$

Ex. $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^T(x_i - \bar{x})$

MLE for Σ if $\underline{X}_1, \dots, \underline{X}_n$ iid $\sim N_p(\mu, \Sigma)$

Estimator of λ_1 where $\boldsymbol{\Sigma} = \sum \lambda_i \underline{e}_i \underline{e}_i^T$?

$\hat{\lambda}_1$ is s.t. $\hat{\Sigma} = \sum \hat{\lambda}_i \hat{\underline{e}}_i \hat{\underline{e}}_i^T$.

□

Distribution of $\bar{\underline{X}}$ and S or $\hat{\Sigma}$

Assume $\underline{X}_1, \dots, \underline{X}_n$ iid $\sim N_p(\mu, \Sigma)$.

Prop. $\bar{\underline{X}} \sim N_p\left(\mu, \frac{1}{n}\Sigma\right)$

proof.

$\tilde{\underline{X}} = \begin{pmatrix} \underline{X}_1 \\ \underline{X}_2 \\ \vdots \\ \underline{X}_n \end{pmatrix} \in \mathbb{R}^{np}$ $np \times 1$ matrix

we're not taking the transpose, this is just a long vector

since each $\underline{X}_i$ is gaussian and they're all independent

$\tilde{\underline{X}} \sim N_{np}\left(\begin{pmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{pmatrix}, \begin{pmatrix} \Sigma & & 0 \\ & \Sigma & \\ 0 & & \Sigma \end{pmatrix} \right)$

$A = \underbrace{\left[I_{p\times p} \; I \; \cdots \; I \right]}_{n} \quad p \times np$

$= \left[\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \\ & & \ddots & \\ 0 & & & 1 \end{pmatrix} \cdots \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \\ & & \ddots & \\ 0 & & & 1 \end{pmatrix} \right]$

$\frac{1}{n} A \tilde{\underline{X}} = \frac{1}{n} \begin{pmatrix} \sum_{i=1}^{n} x_{i1} \\ \sum_{i=1}^{n} x_{i2} \\ \vdots \\ \sum_{i=1}^{n} x_{ip} \end{pmatrix} = \bar{\underline{X}} = \frac{1}{n}\sum_{i=1}^{n} \underline{X}_i$ (p ?)

$\bar{\underline{X}} = \frac{1}{n} A \tilde{\underline{X}} \sim N_p\left(\frac{1}{n} A \begin{pmatrix} \mu \\ \vdots \\ \mu \end{pmatrix}, \frac{1}{n^2} A \begin{pmatrix} \Sigma & & 0 \\ & \ddots & \\ 0 & & \Sigma \end{pmatrix} A^T \right)$

$\sim N_p\left(\mu, \frac{1}{n^2} n\Sigma \right)$

$\sim N_p\left(\mu, \frac{1}{n}\Sigma \right)$

The sample mean ($\bar{\underline{X}}$) is a linear transformation of a big gaussian vector, so it must be gaussian

□

? dist. of S. (Distribution of a random **matrix**)

Def. Let $\underline{Z}_1, \dots, \underline{Z}_m$ iid $\sim N_p(\underline{0}, \Sigma)$ with $\det(\Sigma) > 0$ (Σ $p \times p$)

p×p matrix

Then $\sum_{i=1}^{m} \underline{z}_i \underline{z}_i^T \sim \text{Wishart}(\Sigma, m)$

in principle the parameters are 3: p, Σ, m (we don't stress p since it's already contained in Σ)

(1928 Wishart).

Properties of Wish. dist.

1. $A_1 \sim \text{Wish}(\Sigma, m_1)$, $A_2 \sim \text{Wish}(\Sigma, m_2)$

$A_1 \perp\!\!\!\perp A_2$ stoch indep.

$\Rightarrow$ $A_1 + A_2 \sim \text{Wish}(\Sigma, m_1+m_2)$ — sum of Wisharts

proof.

$A_1 = \sum_{i=1}^{m_1} \underline{z}_i \underline{z}_i^T$ $\quad \underline{z}_1, \dots \underline{z}_{m_1}$ iid $\sim N_p(\underline{0}, \Sigma)$

$A_2 = \sum_{i=1}^{m_2} \tilde{\underline{z}}_i \tilde{\underline{z}}_i^T$ $\quad \tilde{\underline{z}}_1, \dots, \tilde{\underline{z}}_{m_2}$ iid $\sim N_p(\underline{0}, \Sigma)$

$\underline{z}_1, \dots, \underline{z}_{m_1}, \tilde{\underline{z}}_1, \dots, \tilde{\underline{z}}_{m_2}$

$\| \quad \| \qquad \| \qquad \|$

$\underline{w}_1 \; \underline{w}_2 \quad \underline{w}_{m_1}, \underline{w}_{m_1+1}, \dots \underline{w}_{m_1+m_2}$

$\underline{w}_1, \dots, \underline{w}_{m_1+m_2}$ iid $\sim N_p(\underline{0}, \Sigma)$

$A_1 + A_2 = \sum_{i=1}^{m_1} \underline{z}_i \underline{z}_i^T + \sum_{i=1}^{m_2} \tilde{\underline{z}}_i \tilde{\underline{z}}_i^T =$

$= \sum_{i=1}^{m_1+m_2} \underline{w}_i \underline{w}_i^T \sim \text{Wish}(\Sigma, m_1+m_2)$ ∎

2. C $k \times p$ const., $A \sim \text{Wish}(\Sigma, m)$

$\Rightarrow$ $CAC^T \sim \text{Wish}(C\Sigma C^T, m)$ — matrix · Wishart

proof. $A = \sum_{i=1}^{m} \underline{z}_i \underline{z}_i^T$ $\quad \underline{z}_i$ iid $\sim N_p(\underline{0}, \Sigma)$

$CAC^T = \sum_{i=1}^{m} \underbrace{C\underline{z}_i}_{\underline{w}_i} \underline{z}_i^T C^T$ $\quad \underline{w}_i$ iid $\sim N_k(\underline{0}, C\Sigma C^T)$

$\sim \text{Wish}(C\Sigma C^T, m)$. (by def.) ∎

3. $\sigma^2 > 0$, $A \sim \text{Wish}(\Sigma, m)$

$\Rightarrow$ $\sigma^2 A \sim \text{Wish}(\sigma^2\Sigma, m)$. — scalar · Wishart

proof. $\sigma^2 A = \sum_{i=1}^{m} \underbrace{\sigma \underline{z}_i}_{\underline{w}_i} \underline{z}_i^T \sigma$, $\quad \underline{z}_i$ iid $\sim N_p(\underline{0}, \Sigma)$

$\underline{w}_i$ iid $\sim N_p(\underline{0}, \sigma^2\Sigma)$

$\sim \text{Wish}(\sigma^2\Sigma, m)$ (by def.).

4. $A \sim \text{Wish}(\Sigma, m)$ s.t. $\Sigma_{1\times 1} \; \Sigma = [\sigma^2]$ $\quad (p=1)$ ← Wishart defined on 1 dimensional space

$A = \sum_{i=1}^{m} z_i z_i$ $\quad z_1, \dots, z_m$ iid $\sim N_1(0, \sigma^2)$

$\frac{1}{\sigma^2} A = \sum_{i=1}^{m} \underbrace{\frac{z_i}{\sigma}}_{w_i} \frac{z_i}{\sigma}$ $\quad w_1, \dots, w_m$ iid $N_1(0,1)$

$\frac{1}{\sigma^2} A \sim \chi^2(m)$

$A \sim \sigma^2 \chi^2(m)$

it doesn't mean that we take the density of χ^2 and we multiply it by σ^2, it's just an abbreviation for $\frac{1}{\sigma^2} A \sim \chi^2(m)$

$A \sim Wish(\Sigma, m) \quad \Sigma\ 1\times 1$

$\Rightarrow A \sim \Sigma \cdot \chi^2(m)$

The Wishart distribution is a multivariate extension of $\chi^2(n)$

Ex. $0 \neq \underline{c} \in \mathbb{R}^p$, $A \sim Wish(\Sigma, m) \quad \Sigma\ p\times p$

$\underline{c}^T A \underline{c}$?

$\underline{c}^T A \underline{c} \sim Wish(\underline{c}^T \Sigma \underline{c}, m)$ (p2ep 2)

$\underline{c}^T A \underline{c} > 0$ ($\underline{c}^T A \underline{c}^T$ is just a number!)

since it's a number $\Rightarrow$ $Wish(\underline{c}^T \Sigma \underline{c}, m) \sim (\underline{c}^T \Sigma \underline{c}) \chi^2(m)$

$\Rightarrow \underline{c}^T A \underline{c} \sim (\underline{c}^T \Sigma \underline{c}) \chi^2(m)$

meaning $\frac{\underline{c}^T A \underline{c}}{\underline{c}^T \Sigma \underline{c}} \sim \chi^2(m)$ □

Teo $\underline{X}_1, \dots, \underline{X}_m$ iid $\sim N_p(\mu, \Sigma)$

$\Rightarrow \sum_{i=1}^{m} (\underline{X}_i - \bar{\underline{X}})(\underline{X}_i - \bar{\underline{X}})^T \sim Wish(\Sigma, m-1)$

Coro

$$S = \frac{1}{m-1} \sum_{i=1}^{m} (\underline{X}_i - \bar{\underline{X}})(\underline{X}_i - \bar{\underline{X}}) \sim Wish\left(\frac{1}{m-1}\Sigma, m-1\right)$$

$$\hat{\Sigma} = \frac{1}{m} \sum_{i=1}^{m} (\underline{X}_i - \bar{\underline{X}})(\underline{X}_i - \bar{\underline{X}})^T \sim Wish\left(\frac{1}{m}\Sigma, m-1\right)$$

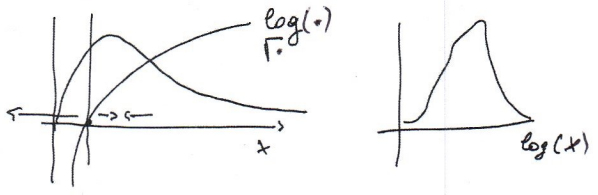
Summary

Teo $\underline{X}_1, \dots, \underline{X}_m$ iid $\sim N_p(\mu, \Sigma)$

1. $\bar{\underline{X}} \sim N_p(\mu, \frac{1}{m}\Sigma)$
2. $(m-1)S \sim Wish(\Sigma, m-1)$
3. $\bar{\underline{X}} \perp\!\!\!\perp S$ (stoch. indepedent)

Teo $\bar{\underline{X}}$ and S are sufficient stat. → if the data is generated by gaussian distribution (no matter μ) then all we need to know is $\bar{\underline{X}}$ and S

Useful transf fr making data "more" Gaussian:



x proportion $\in [0,1]$

$\log \frac{x}{1-x} = \text{Logit}(x)$

LLN

$\underline{X}_1, \dots, \underline{X}_n, \dots$ random vectors i.i.d.
s.t. $E[\underline{X}_1] = \underline{\mu}$, $\text{cov}(\underline{X}_i) = \Sigma$ exist. :

- $\bar{\underline{X}} = \frac{1}{n}\sum_{i=1}^{n} \underline{X}_i \xrightarrow{\mathbb{P}} \underline{\mu}$ as $n \to \infty$
- $S = \frac{1}{n-1}\sum_{i=1}^{n} (\underline{X}_i - \bar{\underline{X}})(\underline{X}_i - \bar{\underline{X}})^T \xrightarrow{\mathbb{P}} \Sigma$ as $n \to \infty$

CLT

$\underline{X}_1, \dots, \underline{X}_n, \dots$ e.v. iid
s.t. $E[\underline{X}_1] = \underline{\mu}$ and $\text{cov}(\underline{X}_1) = \Sigma$ exist
then $\sqrt{n}\,(\bar{\underline{X}} - \underline{\mu}) \sim AN_p(\underline{0}, \Sigma)$

ASINTOTICALLY NORMAL

meaning:
for large n one can approximate the dist. of $\sqrt{n}\,(\bar{\underline{X}} - \underline{\mu})$ with a $N_p(\underline{0}, \Sigma)$

In practice: for large n
$\bar{\underline{X}} \sim N_p(\underline{\mu}, \frac{1}{n}\Sigma)$ approximately.

! It doesn't mean that if the sample is large then it is gaussian. It means that if the sample is large then the sample mean is gaussian