PUBLICATIONS MATHÉMATIQUES ET INFORMATIQUES DE RENNES

J. LEGOUPIL

On the Regression of a Random Variable with Respect to a Stochastic Process and, Possibly, p Random Variables

Publications des séminaires de mathématiques et informatique de Rennes, 1967-1968
« Publications des séminaires du département de mathématiques », , exp. n° 4, p. 1-20
<http://www.numdam.org/item?id=PSMIR_1967-1968____A4_0>

ON THE REGRESSION OF A RANDOM VARIABLE WITH RESPECT TO A STOCHASTIC PROCESS AND, POSSIBLY, p RANDOM VARIABLES

by

J. LEGOUPIL

SUMMARY

In the first section, previous publications are reviewed stating the origin of the problem and the mathematical properties of the linear mean square regression of a random variable Y with respect to a random function X(t). In the second section, estimation problems are studied when eigenfunctions and eigenvalues of the covariance of X(t) are known and when they are unknown by using limit theorems. In section 3, the problem of the choice between the limit theorems is studied, the size n of the sample being given. In section 4, we consider a generalization the case where there are p random variables Z_j and we have to study the multiple regression Y_M of Y with respect to the p, Z_j and the random function X(t).

INTRODUCTION

I began working on this subject in 1958 after a conversation with Pr. Darmois and Pr. Dugué, they told me it would be interesting to think again about the old paper of Fisher "on the influence of rainfall on the yield of wheat at Rothamsted". This paper is in fact a statistical study of the influence of a stochastic process X(t) representing the rainfall during the year T preceding the crop on the random variable Y representing the crop.

Their impressions were that a generalization to stochastic process of the polynomial approximation Weierstrass theorem was implicitely used in the Fisher's paper.

(Dugué 1958 préface page IV and p. 125-138).

With the development of the theory of stochastic processes, it seemed interesting to give a more sound mathematical foundation of Fisher's paper ans hence of the statistical problem of the influence of a stochastic process X(t) on a random variable Y.

Just a few words about the Fisher's paper :

1°) The first method being tried by Fisher was to divide the year in sixty one periods of six days $T_i (i = 1...61)$, X(t) dt being the rainfall during t , t + dt. then he studied the multiple regression of Y with respect to the 61 random variables $\int_{T_i} X(t)\,dt$ rainfall during T_i. But he had only samples of size 65 and it was not possible to get good estimates of the 61 regression parameters.

2°) The method finally used by Sir R. Fisher was to consider a

regression of the form $\alpha_o + \int_T \rho(t)\, X(t)\, dt$ (T = 366 days = 6×61). He thought it was possible to find a good approximation of X(t) with $\Sigma\, A_i\, \rho_i(t)$ $(i = 1,\ldots 6)$. $\rho_i(t)$ being non random functions chosen to be orthogonals and normed over T, A_i are random variables. He was led to a regression $\alpha_o + \Sigma\, (\int_T \rho_i(t)\, \rho(t)\, dt)\, A_i$, which is a multiple linear regression with respect the A_i or with respect to the $\int_T \rho_i(t)\, X(t)\, dt$.

Of course, there were others factors having an influence on the yield of wheat, but Fisher thought that they were not so important and it was possible to remove these effects with appropriate corrections.

Hence the importance of developing a theory relative to the following points :

1) It is necessary to give a precise definition and to set up properties of the mean square linear regression Y_x and of the correlation coefficient C of Y relative to X(t) defined over T.

About approximations, it is necessary to get a rigorous mathematical foundation, justifying the pratical use of such approximations, particulary :

a) Concerning the first method finally not used by Fisher, but that could be useful in other applications, we have to study limit theorems giving Y_x and C as limits, when $L \to +\infty$, of the multiple regression and of the multiple correlation coefficient of Y with the L random $\int_{T_i} X(t)\, dt$ $(i = 1,\ldots L)$ (The T_i being a partition of the year T) every T_i tending to 0.

b) Concerning the second method finally used by Fisher we shall not suppose $X(t) \sim \Sigma\, A_i\, \rho_i(t)$, $(i = 1,\ldots 6)$, we shall study limit theorems

giving Y_x and C as limits when $L \to +\infty$ of the multiple regression and of the multiple correlation coefficient of Y with the L random variables

$$\int_T \rho_i(t)\, X(t)\, dt \qquad (i = 1 \ldots L)$$

Although some specific models have been proposed to describe rainfall (for example Gabriel 1962), in most cases, it does not seem reasonable to admit the validity of a specific model. We have to study statistical problems of estimation and of testing hypotheses using the sample of size n : $Y_1\ X_1(t)$ (observed over T), $Y_2\ X_2(t) \ldots Y_n\ X_n(t)$ (observed over T). We can consider limit theorems when the size n of the sample tends to infinity but in practical applications, often the size of the sample is not very large and we have to think about the possibility of getting some result in the case of not very large samples.

Results about this problem have been presented in 4 "notes aux comptes rendus" (4), J. Legoupil from 1959 to 1962 and an exposition of results obtained on this (and on others related tthings too) up to the end of 1961 has been published in (5). Other results have been obtained since then. Before proceeding further, let us give a brief account of some of the main results in this paper.

1.1. Mean square linear regression.

We make the following assumptions (Hypotheses A) : X(t) is a mesurable stochastic process. $E[X^2(t)]$ exists and is L-integrable over T, Y is a random variable such that $E[Y^2]$ exists.

We study the Infimum of $E(Y - \alpha_o - \int_T \rho(t)\, X(t)\, dt)^2$ for α_o real, $\rho(t) \in L_2(T)$. Let λ_i and $\Psi_i(t)$ the eigenvalues and the ortogonal normed

eigenfunctions of cov $[X(t)\ ,\ X(t')]$.

$$\left[\int_T \text{cov}\ \{X(t)\ X(t')\}\ \Psi_i(t')\ dt' = \frac{\Psi_i(t)}{\lambda_i}\right]$$

Let f_i the coefficients of the development of cov $[Y,X(t)]$ with respect to the $\Psi_i(t)$ (it can be proved that such a development exists, even when the orthogonal normed set of functions $\Psi_i(t)$ is not complete. The convergence of this development is a mean square convergence over T.

$$\text{Let } Y_x = E(Y) + \Sigma\ \lambda_i\ f_i \int_T \Psi_i(t)\ [X(t) - E\{X(t)\}]\ dt$$

We have the following results :

Theorem 1.1.

Under the hypotheses A :

1°) $E[Y - \alpha_o - \int_T \rho(t)\ X(t)\ dt]^2 \geq E[Y - Y_x]^2 = \text{var}\ Y - \sum_i \lambda_i\ f_i^2$

2°) When the series $\sum_i \lambda_i^2\ f_i^2$ converges, there exist α_o and $\rho(t)$ such that the equality in 1) holds, this $\rho(t)$ is a solution of

$$\int_T \text{cov}\ \{X(t),X(t')\}\ \rho(t')\ dt' = \text{cov}[Y,X(t)]$$

(almost everywhere) and

$$\alpha_o = E(Y) - \int_T \rho(t)\ E\ \{X(t)\}\ dt$$

and for these α_o, $\rho(t)$: $\alpha_o + \int_T \rho(t)\ X(t)\ dt = Y_x$.

3°) Even when $\Sigma\ \lambda_i^2\ f_i^2$ does not converge, for every sequence of α_{on} and of $\rho_n(t) \in L_2(t)$ such that :

$$E\ \{Y - \alpha_{on} - \int_T \rho_n(t)\ X(t)\ dt\}^2 \longrightarrow E(Y - Y_x)^2 \quad \text{when } n \to +\infty$$

then :

$$\alpha_o + \int_T \rho_n(t)\ X(t)\ dt \quad \text{mean square} \longrightarrow Y_x.$$

The preceding theorem justifies the following definition :

Y_X is the mean square linear regression of Y with respect to X(t) over T.

Theorem 1.2.

Under the hypotheses A :

a) $E(Y) = E(Y_X)$

b) $\text{cov}\,[Y, Y_X] = \text{var}\, Y_X$

c) $\text{cov}\,[(Y - Y_X), X(t)] = 0$ almost everywhere.

1.2. - Correlation coefficient.

Under the hypotheses A, let C the supremum of the correlation coefficient of Y with $\alpha_o + \int_T \rho(t)\, X(t)\, dt$ for α_o real and $\rho(t) \in L_2(t)$, then

$$C = (\text{var}\, Y_X \,/\, \text{var}\, Y)^{1/2}$$

and

$$\text{var}\,(Y - Y_X) = (1 - C^2)\, \text{var}\, Y$$

1.3. Limit theorems.

Considering the regression Y_X as a limit of a multiple regression of Y with respect to L random variables when $L \to +\infty$, we get the following results :

Theorem 1.4.

Under the hypotheses A, if the interval T is divided in L intervals T_i, the multiple regression Y_L of Y with respect to the L random variables $\int_{T_i} X(t)\, dt$ mean square converges to Y_X when $L \to +\infty$, all the T_i tending to 0.

Theorem 1.5.

Under the hypotheses A, if X(t) is mean square continuous over T, then Y_x is mean square limit of the multiple regression of Y with respect to the L random variables $X(t_i)$ $(i = 1,2...L)$ when $L \to +\infty$ with $t_{i-1} < t_i$, for every i all the $t_i - t_{i-1}$ tending to 0.

Theorem 1.6.

Under the hypotheses A, if $\rho_1(t)...\rho_l(t)...$ is a complete system of functions base of $L_2(T)$, the multiple regression Y_L of Y with respect to the L random variables $\int_T \rho_l(t)\, X(t)\, dt$ $(l = 1,...L)$ mean square tends to Y_x when $L \to +\infty$.

Similar limit theorems exist about the correlation coefficient.

2 - Statistical inference on the regression of Y with respect to X(t)

2.1. Function of regression.

When expectations, variances, covariance and intercovariance functions are known, the linear mean square regression Y_x is known (from what has been mentioned before). But generally such a knowledge is not available.

In favourable situations, we have a good knowledge of the stochastic process (we know $E[X(t)]$ and $\mathrm{cov}[X(t),X(t')]$, and we have to make a statistical study of the influence of this stochastic process on Y. But in most cases, very little is known about the stochastic process.

Suppose we have observed a sample of n independent realizations $X_1(t)$, Y_1, $X_2(t)$, $Y_2 \ldots X_n(t)$, Y_n.

The most natural idea is to try to estimate α_o and $\rho(t)$, when they exist, such that :

$$\alpha_o + \int_T \rho(t)\, X(t)\, dt = Y_X.$$

In that case $\rho(t)$ is solution of the integral equation :

$$\int_T \mathrm{cov}[X(t),X(t')]\rho(t')\, dt' = \mathrm{cov}[Y,X(t)] \qquad \text{a.e.}$$

Let $\mathrm{cov}[X(t),X(t')] = \Gamma(t,t')$ and $\mathrm{cov}\,[Y,X(t)] = \Psi(t)$ for abbreviating notations. We may try to use estimates $\Gamma_n^X(t,t')$ and $\psi_n^X(t)$ of $\Gamma(t,t')$ and $\Psi(t)$ for estimating $\rho(t)$ by $\rho_n^X(t)$ solution of :

$$\int_T \Gamma_n^X(t,t')\, \rho_n^X(t')\, dt' = \Psi_n^X(t) \qquad \text{a.e.}$$

There are difficulties in doing so for the following reasons :

1°) The existence of $\rho(t)$ such that :

$$\alpha_o + \int_T \rho(t)\, X(t)\, dt = Y_X$$

requires the convergence of $\sum_i \lambda_i^2 f_i^2$ and this is not easy to verify. Tests this hypothese (i.e. convergence) are also difficult.

2°) Even when $\sum_i \lambda_i^2 f_i^2$ converges :

a) the function of regression $\rho(t)$ is not necessarily unique

b) When α_o, $\rho(t)$ are such that :

$$\alpha_o + \int_T \rho(t)\, X(t)\, dt = Y_X$$

If we consider α_{o1}, $\rho_1(t)$ such that :

$$E(Y - \alpha_{o1} - \int_T \rho_1(t)\, X(t)\, dt)^2 \text{ is close to } E\left[Y - \alpha_o - \int_T \rho(t)X(t)dt\right]^2$$

generally $\rho_1(t)$ is not close to $\rho(t)$ and this is a major difficulty for a statistical sutdy.

3°) Even when $\sum_i \lambda_i^2 f_i^2$ converges, $\rho(t)$ is unique, and "good" consistent estimates $\Gamma_n^x(t\ t')$, $\varphi_n^x(t)$ exist, the convergence of $\Gamma_n^x(t\ t')$ and $\varphi_n^x(t)$ toward $\Gamma(t\ t')$ and $\varphi(t)$ does not have, as a consequence, the convergence of $\rho_n^x(t)$ to $\rho(t)$. There is generally no such convergence.

In a most favourable case when $\Gamma(t\ t')$ is known, the convergence of $\rho_n^x(t)$ toward $\rho(t)$ is not a consequence of the convergence of $\varphi_n^x(t)$ to $\varphi(t)$. $\{\int_T \Gamma(t\ t')\ \rho_n^x(t')\ dt' = \varphi_n^x(t)\}$. In this case, it is possible to handle the situation. Let $f_{n\nu}$ the coefficients of the development of an estimate $\varphi_n^x(t') \in L_2(T)$ of $\varphi(t')$ with respect to the $\Psi_\nu(t')$. When we have a mean square convergence of $\varphi_n^x(t')$ to $\varphi(t')$, we have $\sum_\nu (f_{n\nu} - f_\nu)^2 \longrightarrow 0$ when $n \to +\infty$. The existence of $\rho_n^x(t')$ solution of the integral equation requires the convergence of $\sum_\nu \lambda_\nu^2 f_{n\nu}^2$ which is not a consequence of our hypotheses and of the consistency of the estimate $\varphi_n^x(t')$. (for we have $\lambda_\nu \to +\infty$ when $\nu \longrightarrow +\infty$). But it is possible to modify an estimate of $\varphi(t')$ for example the following :

$$\varphi_n^x(t') = \frac{1}{n-1}\left[\sum_K (Y_k - \bar{Y})\ \{X_k(t') - \bar{X}(t')\}\right] \qquad (k = 1 \ldots n)$$

with

$$\bar{Y} = (Y_1 + \ldots + Y_n)/n \quad , \quad \overline{X(t)} = \{X_1(t) + \ldots + X_n(t)\}/n$$

such that all the $f_{n\nu} = 0$ for $\nu > \nu_0(n)$ (ν_0 being an increasing function of n).

The mean square convergence over T of $\rho_n^x(t)$ toward $\rho(t)$ requires $\sum_\nu \lambda_\nu^2 (f_{n\nu} - f_\nu)^2 \longrightarrow 0$ when $n \longrightarrow +\infty$, and it is possible to get this

with such a modified estimate of $\varphi(t')$, $\sum_\nu \lambda_\nu^2 f_\nu^2$ being assumed convergent for the existence of $\rho(t)$. Therefore it is possible to handle the situation. But the case of a random process $X(t)$ pretty well known is rare in practical applications. We see how much we have to be cautions for the estimation of $\rho(t)$ even in most favourable cases.

2.2 - Estimation of Y_X

For all the reasons stated before, instead of estimating a regression function $\rho(t)$, we are going to estimate the random variable Y_X defining the regression, whose existence and uniqueness does not require strong assumptions. In order to do this, we can use limit theorems given before : Y_X is the mean square limit of multiple regression Y_L of Y with respect to L random variables when $L \longrightarrow +\infty$. So we have, as soon as $L > L_o(\varepsilon)$, $E\left[Y_X - Y_L\right]^2 < \varepsilon$ ($\varepsilon > 0$ arbitrary small). We are going to estimate the approximation Y_L, the advantage of this is that we have just a finite number of parameters to estimate. For example we shall use the limit theorem 1.6. given before. The results are somewhat different when expectations of Y and $X(t)$ are known and when they are not known. We shall state results when they are unknown. Y_L is defined by the values of α_o α_l ($l = 1...L$) such that :

$$E\left[Y - \alpha_o - \sum_l \alpha_l \int_T \rho_l(t)\, X(t)\, dt\right]^2 \text{ is minimum } (l = 1...L)$$

We consider mean square estimates of α_o α_l , α_o^x α_l^x such that :

$$\sum_i \left[Y_i - \alpha_o - \sum_l \alpha_l \int_T \rho_l(t)\, X_i(t)\, dt\right]^2$$

$(i = 1...n \; ; \; l = 1...L)$

is minimum (n is the size of the sample, Y_i $X_i(t)$ are the observed realizations of Y and X(t)).

Let $X_{il} = \int_T \rho_l(t)\, X_i(t)\, dt$. These estimates α_o^X α_l^X are solution of the following system :

$$\alpha_o^X + \sum_l \alpha_l^X \Big(\sum_{i=1}^{n} X_{il}/n \Big) = \Big(\sum_i Y_i \Big)/n$$

$$\alpha_o^X \Big(\sum_i X_{im}/n \Big) + \sum_l \alpha_l^X \Big(\sum_{i=1}^{n} X_{il} X_{im}/n \Big) = \sum_i \frac{Y_i\, X_{im}}{n}$$

$$(i = 1...n \; ; \; l = 1...L \; ; \; m = 1,2..L)$$

Let E_L the vectorial space having for base $\rho_1(t)...\rho_L(t)$. If there is no function $\rho(t) \in E_L$ such that $\int_T \rho(t)\, X(t)\, dt = 1$ almost surely, the probability of the uniqueness for all $n > N$ of α_o^X α_1^X ... $\alpha_L^X \longrightarrow 1$ when $N \longrightarrow +\infty$. Then :

$$Y_L^X = \alpha_o^X + \sum_{l=1}^{L} \alpha_l^X \int_T \rho_l(t)\, X(t)\, dt$$

almost surely tends to Y_L when $n \longrightarrow +\infty$.

We can study the tendancy of $\sqrt{n}\ \big[\sum_l (\alpha_l^X - \alpha_l)\, \rho_l(t)\big]$ $(l = 1...L)$ and $\sqrt{n}\ (\alpha_o^X - \alpha_o)$ toward gaussian multivariate law, but this does not have as a consequence the tendancy of $\sqrt{n}\ (Y_L^X - Y_L)$ toward gaussian law.

3 - Choice of a limit theorem.

After having reviewed previous publications on this topic we shall speak of further developments. In 1963, A. Zinger asked me the following questions : wich limit theorem to choose and when we choose the theorem 1.6., wich set of complete function $\rho_1(t)...\rho_l(t)...$ to use in a practical application.

First let us say just a few words about cases where the knowledge we have or the hypotheses it is reasonable to assume on Y and X(t) in the field of applications we are considering leads naturally to a choice.

1°) If we study fatigue of a part of an aircraft under random vibrating strength X(t), we can consider one or several random variables Y describing the state of the considered part at the end of the period T the plane is supposed to fly before this part is to be checked or replaced. The value of the random variable Y has to be still safe at that moment. We have to study the influence of X(t) during T on Y. It seems advisable in this case to get the highest resonance frequencies and to use the limit theorem VI, using the $\rho_1(t)$ corresponding to these frequencies.

2°) When we study the influence of a seasonal phenomenon X(t) defined for example on one year on Y (this happens in economics), when it seems reasonable to assume Y is not much affected by short periodical fluctuations of X(t), we may use the theorem 1.6. too with $\rho_1(t)$ sinusoïdal functions with periods T, T/2, T/3. In many economical problems a major difficulty is to get a sample of n independant identically distributed realizations.

3°) When in another field, it seems reasonable to assume that X(t) is a function with independant increments, and that the influences of theses increments on Y are additive, it seems advisable to use limit theorem 1.4. and to consider the mean square linear regression of Y with respect to $X(t_1)..X(t_n)$, or what is equivalent the regression of Y with respect to $X(t_1)$ and the independant increments $X(t_2) - X(t_1)...X(t_n)-X(t_{n-1})$

$t_1 < t_2 \ldots < t_n \quad (\in T)$

We shall now examine the case when there is no previous knowledge leading to a choice.

In every method we consider Y_L multiple regression requiring only the knowledge of C parameters such that :

$$E(Y_L - Y_X)^2 < \varepsilon \qquad (\varepsilon \text{ not too large})$$

as soon as $L > L_0(\varepsilon)$.

Naturally it is not possible when we don't know pretty well Y and X(t) to give the expression of $L_0(\varepsilon)$.

What we can do is to check or to test the hypotheses of $E(Y - Y_L)^2$ much smaller than $E(Y^2)$. If this is the case, we have the following upper bound for the distance between Y_X and Y_L because we have got

$$E(Y_X - Y_L)^2 \leq E(Y - Y_L)^2$$

If we wish to improve the precision of our estimation we may try to use another kind of Y_L say Y_L^1, (or to use a non linear regression but we shall not discuss this point to day). It is possible in some cases not to have an improvement possible. (for example if $Y = Y_L + U$, U independant of X(t)). If we use Y_L^1, instead of Y_L (for example by using another set of functions $\rho_1(t)$), it is generally not advisable to increase the number of parameters ($L' > L$) the increased precisions of Y_L^1, being matched by the decreasing of the quality of estimation of a larger number of parameter when we can't have larger samples.

(When we increase the number of parameters L such that $L \geq n$, we are in the situation when we can adjust perfectly Y_L to fit the data , but

this is not an indication of the quality of the estimation and we have no possibility to test the quality of the fitting).

The best it seems possible to do is to compare estimates $E(Y - Y_L)^2$ for different possible Y_L.

Let us examine the case where

$$Y_L = \alpha_o + \sum_1 \int_T \alpha_1 \rho_1(t) X(t) dt \qquad (1 = 1,\ldots,L-1)$$

We shall use the following :

Theorem 3.1.

Under the hypotheses A the following identity holds for every α_o real $\rho(t) \in L_2(T)$:

$$E\left[Y - \alpha_o - \int_T \rho(t) X(t) dt\right]^2 = E\left[Y - Y_X\right]^2 + E\left[Y_X - \alpha_o - \int_T \rho(t) X(t) dt\right]^2$$

when we try to compare

$$Y_L = \alpha_o + \sum_1 \int_T \alpha_1 \rho_1(t) X(t) dt \qquad (1 = L,\ldots,L-1)$$

and

$$Y_L^1 = \alpha_o + \sum_1 \int_T \alpha_1 \rho_1^1(t) X(t) dt \qquad (1 = 1,\ldots,L'-1)$$

($\rho_1^1(t)$ being another set of functions).

We deduce from the preceding theorem that reducing $E(Y - Y_L)^2$ increases the accuracy of the approximation Y_L of Y_X because from :

$$E(Y - Y_{L'}^1)^2 < E(Y - Y_L)^2$$

résults :

$$E(Y_{L'}^1 - Y_X)^2 < E(Y_L - Y_X)^2$$

Furthemore the increasing of the obtained accuracy (measured by $E(Y_L - Y_X)^2$) is equal to the decreasing of $E(Y - Y_L)$ (when Y_L is replaced by $Y_{L'}^1$)

$$E(Y_L - Y_X)^2 - E(Y_{L'}^1 - Y_X)^2 = E(Y - Y_L)^2 - E(Y - Y_{L'}^1)^2$$

(We have not the same situation about regression functions $\rho(t)$, reducing $E(Y - \alpha_o - \int_T \rho(t)\, X(t)\, dt)^2$ does not have as a consequence $\rho(t)$ get nearer from the regression function $\rho_o(t)$ such that $\alpha_o + \int_T \rho_o(t)\, X(t)\, dt = Y_X$) therefore it seems advisable to divise by random choice the sample of size n, $X_1(t)...X_n(t)$ $Y_1...Y_n$ in two independent completary sets, the first being used for estimating parameters in one or a few reasonable Y_L requiring only the estimation of quite a few number L of parameters. The other part of the sample will be used to get estimates of $E(Y - Y_L)^2$ and so to test the adequacy of what has been obtained by using Y_L and by the estimation using the first subset.

4 - Generalization.

4.1. Introduction.

I have been led to the following generalization from the following problem of strength of materials. Schematically we have p random variables Z_j (j = 1...p) caracterizing the state of a material at the end of the fabrication process. We consider the random strains X(t) during the given time T this material is to be used before being replaced or checked up. One or several random variable(s) Y caracterize the state of the material at the end of the period T. We consider the influence of the Z_j and X(t) (during T) on Y. We have no precise indication on the nature of the stochastic process X(t). We have a finite number n of independent observations realized on Y , X(t) (over T) and the Z_j (Y_1 $X_1(t)$ $Z_{j1}...Y_n$ $X_n(t)$ Z_{jn}). (We meet the same kind of problem when we study the influence on the yield

of wheat Y not only of the rainfall X(t) but of others factors Z_j too).

We shall define and study an optimum linear mean square regression Y_M of Y with respect to the Z_j , X(t) (over T), this being done in order to study statistical estimation.

4.2. Hypotheses B.

X(t) is a measurable process defined over T, $E[X^2(t)]$ exists and is L-integrable over T, Y and the Z_j have second order moments.

4.3. Notations :

Let $\Psi_i(t)$, λ_i the orthogonal normed eigenfunctions and eigenvalues of $\mathrm{cov}[X(t)\ X(t')]$, let f_i and f_{ji} the coefficients of the development of $\mathrm{cov}[Y,X(t)]$ and $\mathrm{cov}[Z_j,X(t)]$ with respect to the $\Psi_i(t)$. (It is possible to prove the existence of this development even when the $\Psi_i(t)$ is not a complete system, the convergence of this development being mean square convergence). Let the α_j^o a solution of the system :

$$\sum_j \alpha_j^o \left[\mathrm{cov}(Z_j\ Z_{j'}) - \sum_i \lambda_i\ f_{ji}\ f_{j'i}\right] = \mathrm{cov}(Y\ Z_{j'}) - \sum_i \lambda_i\ f_i\ f_{j'i}$$

$$(j = 1...p \quad ; \quad j' = 1...p)$$

(there is always a solution, but it is possible to have more than one solution, this fact being connected with the linear dependence or independence of X(t) over T and the Z_j).

Let Δ_p the determinant of the coefficient of the α_j^o in the preceding system and Δ_{p+1} the determinant obtained by adding to Δ_p a last row and a last column whose elements are :

$$\delta_{p+1,j} = \delta_{j,p+1} = \mathrm{cov}(Y\ Z_j) - \sum_i \lambda_i\ f_i\ f_{ji} \qquad (j = 1...p)$$

$$\delta_{p+1,p+1} = \mathrm{var}\ Y - \sum_i \lambda_i\ f_i^2$$

Let $C_i^o = \lambda_i f_i - \sum_j \alpha_j^o f_{ji}$

and

$$Y_M = E(Y) + \sum_j \alpha_j^o [Z_j - E(Z_j)] + \sum_i C_i^o \int_T \Psi_i(t) \left[X(t) - E\{X(t)\}\right] dt$$

(The convergence of $\sum_i$ being m.s. convergence which is a consequence of the hypotheses B).

4.4. mean square regression.

Theorem 4.1.

Under the hypotheses B, Y_M is unique, even when the α_j^o C_i^o are not so.

Theorem 4.2.

Under the hypotheses B :

1°) $\underset{\alpha_o \alpha_j \text{ reals}}{\text{Inf}} E\left[Y - \alpha_o - \sum_j \alpha_j Z_j - \int_T \rho(t) X(t)\, dt\right]^2 = E(Y - Y_M)^2$

$\rho(t) \in L_2(T)$.

2°) If and only if the series $\Sigma \lambda_i^2 f_i^2$ and $\sum_i \lambda_i^2 f_{ji}^2$ $(j = 1 \ldots p)$ converge, there exist α_o α_j and $\rho(t) \in L_2(T)$ such that :

$$E(Y - \alpha_o - \Sigma \alpha_j Z_j - \int_T \rho(t) X(t)\, dt)^2 = E(Y - Y_M)^2$$

and for these α_o , α_j , $\rho(t)$:

$$\alpha_o + \Sigma \alpha_j Z_j + \int_T \rho(t) X(t)\, dt = Y_M$$

Theorem 4.3.

Under the hypotheses B, we have :

$$E\left[Y - \alpha_o - \sum_j \alpha_j Z_j - \int_T \rho(t) X(t)\, dt\right]^2 =$$

$$= \text{var}(Y - Y_M) + E\left[Y_M - \alpha_o - \Sigma \alpha_j Z_j - \int_T \rho(t) X(t)\, dt\right]^2$$

$\forall\, \alpha_o$, α_j reals and $\rho(t) \in L_2(t)$.

This theorem is fundamental for application of a method analogous to the end of section 3.

Theorem 4.4.

If we assume hypotheses B and if $\Delta_p \neq 0$,

$$E(Y - Y_M)^2 = \Delta_{p+1} / \Delta_p .$$

Theorem 4.5.

Under the hypotheses B, we have :

$$E(y) = E(Y_M) \quad ; \quad \operatorname{cov}(Y\ Y_M) = \operatorname{var} Y_M$$

$$\operatorname{cov}\{(Y - Y_M)\ X(t)\} = 0 \quad \text{a.e.} \quad \operatorname{cov}\{(Y - Y_M)\ Z_j\} = 0 \qquad \forall\ j.$$

Theorem 4.6.

Let : $C = \sup$ correl. coeff. $\left[Y\ ,\ \alpha_o + \Sigma\ \alpha_j\ Z_j + \int_T \rho(t)\ X(t)\ dt\right]$ for $\alpha_o\ \ \alpha_j$ reals $\rho(t) \in L_2(T)$. Under the hypotheses B, we have :

1°) $C = (\operatorname{var} Y_M / \operatorname{var} Y)^{1/2}$

2°) $\operatorname{var}(Y - Y_M) = (1 - C^2) \operatorname{var} Y.$

4.5. Some indications on statistical inference

The preceding properties of Y_M have been stated in order to make statistical inference. The knowledge of Y_M requires the knowledge of an infinite number of parameters. When eigenfunctions of $\operatorname{cov}[X(t)\ X(t')]$ are known, it is possible to study estimation problemof a regression function generalizing what has been studied at the beginning of section 2. When they are unknown, it is possible to prove limite theorems generalizing theorems 1.4., 1.5., 1.6., and to study limit theorems about mean square estimation of an approximation Y_L of Y_M when the size n of the sample

tends to infinity along the same lines as in section 2.

When we don't know much about the stochastic process and when the size N of the sample is given we have the same kind of problem as in section 3. For example we may, from a random subsample of size N_1, decide which of the Z_j and which of the functions $\rho_i(t)$ seems to have the most significant influence on Y and to use the other results (subsample of size $N - N_1$) in order to test the adequacy of what has been got.

In many applications $\rho_i(t)$ will be either trigonometric functions of period T/i (fatigue problems for example). In other application an exponential increasing function $\rho(t)$ can be used and we have to make an optimum choice of the number L of parameters (and of the Z_j to use) in the approximation Y_L of Y_M. Clearly when L is too large we can have Y_L close to Y_M but Y_L^x estimate of Y_L is bad (the size N of the sample is not too large). On the opposit when L is too small we can get good estimate Y_L^x of Y_L but Y_L is not close to Y_M estimate Y_L^x of Y_M.

We can proceed along the same lines as in section 3 in order to get and to test the adequacy of the estimate Y_L^x of Y_M.

BIBLIOGRAPHY

(1) R.A. FISHER : The influence of rainfall on the yield of wheat at Rothamstead (Philo. Trans. of the Royal Stat. So. London Series B, vol. 213, p. 89-142).

(2) D. DUGUE : Traité de statistique théorique et appliquée, Paris (Masson 1958).

(3) GABRIEL and NEUMANN : A Markov chain model for daily rainfall occurence at Tell Aviv (Quart. Journal Royal meteorological So. 88 , 90-5 , 1962)

(4) J. LEGOUPIL : 4 notes , Comptes rendus aca. Sci. Paris, T. 249, p.1444 T. 254, p. 621 , T. 255, p. 634 , T. 255, p. 2903 from 1959 to 1962 announcing results developed in the following.

(5) J. LEGOUPIL : Sur la regression, la corrélation et la dépendance de probabilité d'une variable aléatoire par rapport à une fonction aléatoire (Public. Inst. Stat. Uni. Paris, Fasc. 3 et 4, p. 165-314) 1963.

(6) J. LEGOUPIL : Advances in the statistical study of a finite set of random variables and a stochastic process by regression methods. Contributed paper presented at the European meeting of Statisticians, London Sept. 1966.