STATISTICAL INTERACTION REVISITED

by

Regina C. Elandt-Johnson

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1457

March 1984

STATISTICAL INTERACTION REVISITED

Regina C. Elandt-Johnson
Department of Biostatistics, University
of North Carolina, Chapel Hill, N.C. 27514

Abstract

To have valid meaning, "statistical interaction" must be consistent with biological or physical interaction which is defined as mutual (dependent) action of factors contributing to the response effect. In mathematical and statistical response models the parameter representing interaction depends on the model concerned. Factorial experiments, discrete response, and regression models are discussed. The commonly used definition of statistical interaction -- as "departure from additivity" -- has valid meaning only, where response is represented as a sum of contributions of different factors, and cannot be accepted as the definition of interaction.

Key Words: Factor; Response; Factorial experiment; Effect modifying factor; Multiplicative model; Contingency table; Regression; Hazard rate.

This work was supported by U.S. National Heart. Lung, and Blood Institute contract NIH-NHLI-712243 from the National Institutes of Health.

1. INTRODUCTION

1.1. There is a consensus (though never "legally" established) among statisticians that statistical interaction is particularly associated with additive models -- no matter what they represent -- and it is often defined as "departure from additivity." Consequently, it is then concluded that models which are not additive must include interaction. This, of course, confuses many biologists, epidemiologists and also statisticians in situations where it is obvious that the factors act, for example, in a multiplicative, but independent fashion.

To get around this difficulty, some authors speak about biological interaction as contrasted with statistical interaction (Rothman *et al.* (1980)); about model dependent interaction (Kupper and Hogan (1978), Walter and Holford (1978)); about additive and multiplicative interaction (Darroch (1974), Kleinbaum *et al.* (1982)); and some other kinds of interactions such as public health interaction (Rothman *et al.* (1980)), causal interaction (Koopman (1981)), etc. This stratagem does not satisfy many scientists, especially researchers in epidemiology, where a correct and unique meaning of interaction is especially important in studying the contributions of different risk factors to disease occurrence and mortality (Rothman (1978), Koopman (1981)). The quotation below is from Rothman (1978).

"...Suppose two identical sets of data were given to two statisticians in different universities. One statistician, examining the data for biologic interaction, uses a local computer program which is based on the premise that an additive model describes independent effects. The other statistician uses a different program based on a log-linear model. Because the criterion

used to assess interaction is different for these two data analysts, it is conceivable that one could conclude that the data demonstrate interaction while the other concludes opposite... Logic, however, dictates that a state of nature cannot be both present and absent simultaneously... Statistical models which lead in principle to divergent conclusions about the same data cannot be equally acceptable if scientific inference is the goal -- this much should be axiomatic even without agreement as to the actual choice of model... ."

Though Rothman's argument has a point, it would not have arisen, if the definition of "response model" and the meaning of "interaction" were clarified.

1.2. We consider different *factors* (stimulants, variables, events) acting on or causing a phenomenon, or a result, or a reaction called *response*. According to the New Webster's Dictionary (1981), we quote: "Response - the act of responding or replying; *biol.* the activity or behavior of an animal or a plant as a result of stimulation; reaction." In this context, biological meaning of response is most plausible, and the factors play the role of the stimulants.

In this Dictionary we also read: "Interact - to act on each other;" "Interaction - mutual or reciprocal action."

We shall, then, consider models in which response is represented as a function of effects or contributions of different factors associated with response. If the effect (or contribution) of factor A is not influenced by (does not depend on the action of) factor B, and vice-versa, then there is *no interaction*; as we will see later, an additive model is a special case of response model, and departure from additivity is a special case of interaction but *not the* definition of it.

We shall discuss three basic types of response models in which interaction may occur: (i) traditional factorial experiment models; (ii) binary response models and their probabilities (contingency tables); (iii) linear and non-linear regression models.

2. FACTORIAL EXPERIMENTS

Historically, the term "interaction" was used by R.A. Fisher in analysis of variance of factorial agricultural experiments. The response variable was a quantitative characteristic (e.g. yield), whose values were influenced by different factors controlled at several levels or classified into several categories. To the best of my knowledge, there is no formal definition of "statistical interaction" in either R.A. Fisher's or his collaborators' work; the term is used in its linguistic meaning, mentioned in Section 1 of the present paper. For example, in his book "The Design of Experiments" ((1960), first edition (1935)) on p. 94, Fisher says: "... The modifications possible to any complicated apparatus, machine or industrial process must always be considered as potentially *interacting with one another* (my italics), and must be judged by the probable effects of such interactions... ."

A typical response model for agricultural data is given here in Example 1. For simplicity, only two factors, A and B, will be used.

EXAMPLE 1. Yield of wheat (x) depends on variety of wheat (A-classification) and fertilizer combination levels (B-classification). Let $\xi_{ij} = E(X_{ij})$ be the expected yield for the ith variety with the jth fertilizer level.

Assume first that the effects of varieties and the effects of fertilizer combination levels act *independently* on the resulting yield. A reasonable response model, in which yield is the response variable, might be

$$\xi_{ij} = \mu + \alpha_i + \beta_j, \tag{2.1}$$

where μ is the overall mean, α_i is the effect of the ith variety and β_j is the effect of the jth fertilizer combination. This model is called an *additive* response model. It implies that the effect of the ith variety is the same (α_i), whatever the level of the fertilizer combination might be. And similarly, for the jth level of the fertilizer combination. This also implies that for two different varieties, (i) and (i'), the *difference*,

$$\xi_{ij} - \xi_{i'j} = \alpha_i - \alpha_{i'} \quad , \tag{2.2a}$$

is constant for all j; and similarly,

$$\xi_{ij} - \xi_{ij'} = \beta_j - \beta_{j'} \tag{2.2b}$$

is constant for all i.

If, however, the effects of the varieties *depend* (biologically and mathematically) on fertilizer combinations, the response model can be written

$$\xi_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \quad , \tag{2.3}$$

where γ_{ij} represents the *interaction* between varieties and fertilizer levels. In this case, interaction is, indeed, *departure from additivity*. But, equivalently, it can also be expressed in terms of *differences*. For example, the difference (2.2a) takes now the form

$$\xi_{ij} - \xi_{i'j} = (\alpha_i - \alpha_{i'}) + (\gamma_{ij} - \gamma_{i'j}) \quad , \tag{2.4}$$

which, in general, is not the same for all j.

In this example, where factor B not only contributes to the response ξ_{ij}, but also acts as a modifier of the effect of factor A, factor B is called an *effect modifying factor*. The modification in this model is measured in terms of the interaction, γ_{ij}.

Of course, by the same token, A may be formally considered as a modifier of the effect of B, though such a statement seems to be rather odd in this context, in view of the causal relationship of these factors.

EXAMPLE 2. Suppose that the response variable is the income (y) (in dollars) from the wheat crop from J different regions with prices per unit weight depending on I different seasons, but *independent* of the region. If $E(y_{ij}) = \eta_{ij}$, then

$$\eta_{ij} = \omega_i \xi_j, \tag{2.5}$$

where ω_i is the expected price in the ith season, and ξ_j is the expected yield of wheat in the jth region. Model (2.5) is by its nature, *multiplicative*. It is not a biological model, but we still may consider the price and the yield as "factors," and model (2.5) as "response model" without interaction. Here the *ratios*

$$\eta_{ij}/\eta_{i'j} = \omega_i/\omega_{i'} \tag{2.6a}$$

are *constant* for all j, and similarly,

$$\eta_{ij}/\eta_{ij'} = \xi_j/\xi_{j'}, \quad \text{for all i.} \tag{2.6b}$$

If, however, price depends not only on the season but also on the region, then we will have a *multiplicative model with dependent action* (i.e., interaction) of the form

$$\eta_{ij} = \omega_i \xi_j \kappa_{ij} . \tag{2.7}$$

Using logarithmic transformation, models (2.5) and (2.7) become

$$\log \eta_{ij} = \log \omega_i + \log \xi_j \tag{2.8}$$

and

$$\log \eta_{ij} = \log \omega_i + \log \xi_j + \log \kappa_{ij} , \tag{2.9}$$

respectively.

3. DISCRETE RESPONSE MODELS

Of special interest are situations in which the response is binary: -- "yes" or "no" -- and the data are expressed in terms of frequencies, arranged in the form of a contingency table with I × J cells. We now consider *models of probabilities of response* (occurrence of event E equivalent to "yes").

Let p_{ij} (>0) be the probability of occurrence of event E at the ith level of factor A and the jth level of factor B. If the factors A and B act *independently* on the response, then

$$p_{ij} = \delta_i \theta_j \ , \tag{3.1}$$

where δ_i and θ_j are the corresponding marginal probabilities. If, however, the factors do not act in an independent manner, we may have

$$p_{ij} = \delta_i \theta_j \varepsilon_{ij} \ , \tag{3.2}$$

where ε_{ij} represents an interaction effect.

Similar arguments apply when considering incidence *rates* rather than probabilities.

Multiplicative models of the kind (3.1) or (3.2) are of special interest to epidemiologists in studying the effects of various risk factors on disease occurrence and mortality. For illustration, we consider two examples.

EXAMPLE 3. Consider an I × J contingency table in which the event of interest is death in J populations, each stratified into the number I age groups. Death rates are functions of both population effect and age effect and, in this example, we assume that these factors act *independently*.

Then the death rate in the ith age group and the jth population could be expressed by a model analogous to (3.1), that is

$$\lambda_{ij} = \delta_i \theta_j ; \tag{3.3}$$

there is *no interaction* between age and population in their contribution to mortality.

EXAMPLE 4. Asbestos and smoking are both factors for lung cancer. However, lung cancer incidence among people exposed to asbestos is much higher in smokers than in nonsmokers of the same age. A model for death rates from lung cancer,

$$\lambda_{ij} = \delta_i \theta_j \varepsilon_{ij} , \tag{3.4}$$

might be plausible here. The parameter ε_{ij} represents interaction. Clearly, smoking is a factor which stimulates the action of asbestos on the occurrence of lung cancer, resulting in increased cancer incidence rate; thus smoking *acts as an effect modifying factor.*

Incidentally, this example is also discussed by Rothman (1978), who argues that *if* the data were well fitted by the model

$$\lambda_{ij} = \alpha_i + \beta_j , \tag{3.5}$$

then there will be no interaction and no effect modification, which contradicts the reality, since smoking nad asbestos are interactive factors. Though it is conceivable that (3.5) might be fitted to some data, this model would have no biological (or even statistical) interpretation, and would be, in my opinion, inappropriate. Since λ_{ij} is not a response but a *rate of*

response, α_i and β_j should also be rates associated with smoking and asbestos.

4. REGRESSION MODELS

4.1. Linear models. A model linear in the parameters β_i's ,

$$E(y) = \eta = \beta_0 + \beta_1 g_1(\underset{\sim}{x}) + \ldots + \beta_S g_S(\underset{\sim}{x}), \tag{4.1}$$

where $\underset{\sim}{x} = (x_1, \ldots, x_k)$ are k independent (predicting) variables, the $g_s(\underset{\sim}{x})$'s are their functions, and y is the dependent (random) variable, is called *a linear in parameters regression model*. We confine our further discussion to models in which the $g_s(x)$'s are linear functions or are products of linear functions of the x's, and, for simplicity, take k = 2, that is, we have a model

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 \ . \tag{4.2}$$

Model (4.2) can be considered as a response model, in which E(y) is a response, and x_1, x_2 play the roles of factors. The product term, $\beta_{12} x_1 x_2$, respresents the joint contribution (interaction) of x_1 and x_2 to E(y). Estension to more than two independent variables is straightforward.

4.2. Exponential hazard rate functions with concomitant variables. Various time dependent phenomena such as, for example, biochemical reactions, growth, and instantaneous mortality rates (hazard rates) can be represented by exponential functions. We confine ourselves to hazard rate function of the form

$$\mu(x) = Re^{\alpha x}, \quad x>0, \quad R>0, \quad \alpha>0. \tag{4.3}$$

We note that at x=0, $\mu(0) = R = e^{\phi} > 0$, which implies that at birth there is a positive force of mortality for each individual; the parameter R may be interpreted as some initial level of "toxicity." Since R is usually small (of order 10^{-4}), ϕ must be negative. The quantity $-\phi$ (>0) might be interpreted as some initial level of 'vital resources' which will be used up by the living process in which the physiological characteristics (blood pressure, cholesterol, etc.) as well as some external variables (smoking, diet, etc.) play the roles of risk factors. For simplicity, consider only two risk factors, denoted by z_1 and z_2, besides age which is denoted by x. We consider the parameter ϕ as a function of z_1, z_2, and the function

$$\psi(x;z_1,z_2) = \phi(z_1,z_2) + \alpha x \tag{4.3}$$

as an *intrinsic response*. If the z's do not interact, and their effects in (4.3) can be represented by a linear function, then we have

$$\psi(x,z_1,z_2) = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \alpha x \ . \tag{4.4}$$

The model

$$\mu\ (x;z_1,z_2) = \exp(\beta_0 + \beta_1 z_1 + \beta_2 z_2 + \alpha x) \tag{4.5}$$

is *intrinsically linear* (Draper and Smith (1966)), since under (logarithmic) transformation it becomes linear -- it is well known as a *log-linear model*.

Confining the model only to first order interaction terms, we have

$$\mu(x;z_1,z_2) = \exp[(\beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_{12} z_1 z_2) + (\alpha_0 x + \alpha_1 z_1 x + \alpha_2 z_2 x)]. \tag{4.6}$$

The full (saturated) model would include an additional term $\alpha_{12} z_1 z_2 x$.

The hazard rate function in (4.5) is an exponential *function of response* $\psi(x;z_1;z_2)$; it is intrinsically linear and includes interaction terms.

4.3. Additive hazard rate models. The overall (from all causes) hazard rate, $\mu(x)$, can be represented as a function of rates for specific causes. In competing risk theory, it is often assumed that each death is due only to a single cause (i.e., deaths from different causes are mutually exclusive events). Assuming that there are k causes operating simultaneously, we have

$$\mu(x) = \mu_1(x) + \mu_2(x) + \ldots + \mu_k(x), \tag{4.6}$$

where $\mu_i(x)$ is the hazard rate for the ith cause. Of course, in this model $\mu(x)$ plays the role of the response and the $\mu_i(x)$'s -- the roles of factors.

If cause C_i were associated with a 'factor' z_i, and if the rate $\mu_i(x)$ were approximately proportional to z_i, then we would have the additive model

$$\mu(x;\underset{\sim}{z}) \doteq \beta_1 z_1 + \beta_2 z_2 + \ldots + \beta_k z_k + \alpha x . \tag{4.7}$$

It is, however, rather implausible to assume that there is a single risk factor responsible for each cause, and so it would be difficult to interpret model (4.7), though it might, perhaps, be suitable for a short interval of age.

5. DISCUSSION

5.1. Mathematical and statistical models constructed for the purpose of describing biological, physical, economical (or some other) phenomena are useful, if their variables and parameters have real interpretation in regard to the phenomenon they described.

5.2. Mathematical models of interest, in this article, are response models. They can be classified in two broad classes: (a) models with quantitative response (factorial experiments and regression models), and (b) discrete (binary) response models.

5.3. Interaction (biological, physical) of factors contributing to a response is defined as their mutual (dependent) action. If it so happens that the model is represented as a sum of contributing factors, then interactions coincide with the departure from a model in which the effects of the factors are otherwise additive. To the best of my knowledge, there is no claim in R.A. Fisher's work that this is *the* definition of statistical interaction. Statisticians working in the field of factorial experiments and ANOVA used "departure from additivity" as *the* definition of statistical interaction--which for these models is correct. For some (unknown) reasons this definition was generalized to other (not necessarily response) models, and caused some confusion and contradictions.

5.4. Those who still cannot "give up" the use of the definition of statistical interaction as "departure from additivity" might find the paper by Darroch and Speed (1983) on generalized definition of statistical interaction of some interest.

REFERENCES

DARROCH, J.N. (1974), "Multiplicative and Additive Interaction in Contingency Tables," Biometrika, 61, 207-214.

DARROCH, J.N., and SPEED, T.P. (1983), "Additive and Multiplicative Models and Interactions," Annals of Statistics, 11, 724-738.

DRAPER, N.R., and SMITH, H. (1966), Applied Regression Analysis, 1st ed. New York: J. Wiley, Chapter 10.

FISHER, R.A. (1960), The Design of Experiments, 5th ed. New York: Hafner Publishing Company, Chapter 4.

KLEINBAUM, D.G., KUPPER, L.L., and MORGENSTERN, H. (1982), Epidemiologic Research: Principles and Quantitative Methods, Belmont, California: Lifetime Learning Publication, Chapter 19.

KOOPMAN, J.S. (1981), "Interaction Between Discrete Causes," American Journal of Epidemiology, 113, 716-724.

KUPPER, L.L., and HOGAN, M.D. (1978), "Interaction in Epidemiologic Studies," American Journal of Epidemiology, 108, 447-453.

NEW WEBSTER'S DICTIONARY of the English Language (1981), New York; Delair Publishing Company.

ROTHMAN, K.J. (1978), "Occam's Razor Pares the Choice Among Statistical Models," American Journal of Epidemiology, 108, 347-349.

ROTHMAN, K.J., GREENLAND, S., and WALKER, A.M. (1980), "Concepts of Interaction," American Journal of Epidemiology, 112, 467-470.

WALTER, S.D., and HOLFORD, T.R. (1978), "Additive, Multiplicative and Other Models for Disease Risks," American Journal of Epidemiology, 108, 341-346.