*" sometimes the independent variables **interact** so that the effect of one is different depending on the value of one more of the other independent variable variables" (line 6, 1st paragraph, page 94)*

The word "interact" is a less-than-than-helpful term used by statisticians. The correct meaning of interact is, according to

Merriam-Webster's COLLEGIATE® DICTIONARY

interact (intransitive verb) : 1839 : **to act upon one another.**

Merriam-Webster's COLLEGIATE® THESAURUS

interact:	**to act upon one another**
Synonyms	**coact, interplay, interreact**
Related Word	**collaborate, cooperate; combine, join, merge, unite**

Example(osm) : love and time interact: "love makes time pass and time makes love pass".

In the dictionary meeting, there are just two objects: there is **no third object on which** the two act. Contrast this with "*the effect of X1 [**on the response variable Y**] is different depending on the value of X2"* above.

The terms **collaborate, cooperate** in the thesaurus do explain how the meaning has drifted and how some might take interaction as a synonym for "synergy".

The term "modification" describes much better the concept that "*the effect of one X [on the response variable Y] is different depending on the value of another X"*.

We can say that "X2 *modifies* the Y-X1 relationship" and conversely, "X1 *modifies* the Y-X2 relationship". Or, to use Hanley's definition, there are "*different slopes for different folks*".

Another mathematically correct way to describe what statisticians call an "interaction" is *"a regression situation with product terms involving 2 (or more) X variables"*.

" equation is still linear in the regression coefficients b_i "
(2nd paragraph page 94)

Here, "linear" means combinations of the b's, not of the X's.

"solve the resulting multiple regression problem with the three independent variables X1, X2, and X12"

I complained early on about the choice of the term "*independent*" variables. Here you can see why the word independent is a poor choice. Since X12 is a product of X1 and X2, one cannot vary X12 "independently" of X1 and X2.

" indicates that there is a <u>significant interaction between the independent variables</u> i.e. the effect of one depends on the value of the other"

Again it's a little misleading to just say that there is a significant interaction *between the two independent variables,* without even mentioning the response variable! The second part of the sentence is more accurate and also more descriptive.

Equations on p94:

If I were motivating the equations, I would start with the second one. It shows that there is *no single* "Y -on-X2" *slope*. Rather the "Y -on-X2" slope is itself a function of X1

$$\text{"Y -on-X2" slope} = B_2 + B_{12}X_1$$

Likewise, the intercept in the "Y vs. X2" relation is again a function of X1

$$\text{Intercept in "Y vs. X2" relation} = B_0 + B_1X_1$$

So there are as many intercepts and slopes as there are values of X_1.

I would then expand this second equation to get to the first form of the equation mentioned in paragraph 2.

"The sensitivity of $\mu_{y|x}$ to changes in X2 depends on the value of X1"

That's a clear way to say it.

"As a general rule, one should always include the first order terms for all variables included as interaction terms in regression equation"

Whether to do so will depend somewhat on the "biology". It is interesting that in the situations analyzed in this chapter, the authors were able to dispense with the need for one or more of the first order terms in the regression equation.

How bacteria live in salty environments

"At lower salt concentrations, the slopes of the lines are steeper than a higher concentrations ... activity appears to decrease with increasing salt concentrations .. " (middle of page 95)

From this clear verbal description, one then goes to the mathematical representation given in equation 3.21.

" ... we define the new variable I" (for interaction)... " (6 lines from end of p95)

It's much better to say we define a **product term**

"Interesting conclusions ... the intercept term is not significantly different from zero" (2nd paragraph page 97)

One shouldn't rely on significance tests to decide whether to include certain terms in a regression. Here, there's a good *physical* reason why the B_0 should be zero -- and just as good reason why the term B_s should also be zero. Again, rather than beginning with general mathematical models and eliminating terms from them, why not consider the physics of the situation -- which would suggest that the lines form "rays" emanating from zero?. Then, it's a matter of turning these verbal descriptions and graphic descriptions into equations.

" The original question we asked was whether the rate at which tritiated water was produced depended on salt concentration" (bottom of page 98)

This is a clearer way of stating it than asking whether there was an "*interaction*" between time and concentration.

SIMILAR EXAMPLE: the data on the effects of alcohol on the smooth pursuit velocity of the eye (on web page)

SUMMARY

So far, the authors' data examples are mostly from "well controlled isolated physiological experiments" i.e., from basic science. They refer to the situation where "experimenters are not able to independently manipulate each predictor variable". In working with intact human beings as their research material, epidemiologists are seldom able to independently manipulate *any* of the predictor variables: subjects choose their own levels for their X variables, and investigators study them intact.

The second paragraph has a nice summary of some of the reasons for carrying out multiple regression. First, it can be used as a descriptive tool. Second, it can be used to estimate the relative importance of different "explanatory" or predictor variables. Third, it gives quantitative estimates of

the parameters associated with each of these, together with standard errors which measure the reliability of these estimates.

The authors make an **important point that multiple linear regression is not just about straight lines and planes**! The use of square terms and product terms can produce complex curved response surfaces. The equations produced by these terms still fall within the class of models known as multiple linear regression: the equations are still (linear) combinations of the B's and the mathematical *terms* formed from the original variables. (By the way, it would help if more texts made the distinction between the numbers of *variables* and numbers of *terms* in a regression: there can be many more of the latter!)

"Although we have assumed that the dependent variable varies normally about the regression plane, we have not had to make any such assumptions about the distribution of values of the X's" (end of p 103, beginning of p 104)

This is another important point. In regression problems, the X's are treated as if there were **chosen** by the investigator. In a statistical sense, the X's are not random variables.

"... multiple regression rests on a set of assumptions about the <u>population</u> from which the data were drawn " (last para)

Don't think of a single "population" behind the sample. A regression line connects the means of the response variable at different at different X values. There are as many conceptual "(sub)populations" as there are possible X values.

The last paragraph on p 104 summarizes again the elements in, and assumptions behind, a multiple regression:

> *"The mean value of Y for any given set of X's is described by the regression equation"*
>
> *"The variation about the surface of means is Gaussian"*
>
> *"The amplitude of this variation is the same regardless of the values of the X's"*
>
> "we also assume that
> ***the independent variables are statistically independent, i.e., knowing the value of one of more of the independent variables is no information about the others Although this latter condition is rarely strictly true in practice, it is often reasonably satisfied"***

I do not understand why the authors need to "assume" -- or "require" this "independence" **(i.e. absence of correlation among)** the X's. In basic science research, yes, it may be feasible to have "balanced" designs, where the distribution of the values of factor 1 is the same at all values of X2, so that X1 and X2 are *statistically* independent. But in epidemiolgy and other non-experimental research, it is quite unusual to have statistical independence (or even lack of correlation). It is common to have the multi-collinearity they speak of: i.e. correlation between two X's or between one X and some combination of the other X's.

Even with multi-collinearity, multiple regression still provides unbiased estimates. However, it takes more observations to make them as precise as if the X's were statistically uncorrelated. Maybe this is what the authors mean by "*Although this latter condition is rarely strictly true in practice, it is often reasonably satisfied. "*

I would have put it differently:

"These is often a moderate amount of collinearity in the X's; the consequences of multicollinearity are not critical if the collinearity is not extreme. The consequences are also a function of the numbers of observations one has. The same amount of collinearity will have lesser consequences (greater reliability) if the sample size is larger and more consequences (poorer reliability) if the sample size is smaller"

To see this, look again at equation 3.6 (2 X's, p 58) and equation 3.16 (multiple X's, p 79) for the standard error of the estimated regression coefficients.. But rewrite the equations so that the $\sqrt{n}$ is explicit in the denominator of the standard error -- as I do in my notes on pages 51-62 of CHAPTER THREE.