Multiple linear regression (paragraph one, p50 [54])

The word "*linear" in* multiple linear regression

The authors never explicitly defined what they mean by the term "linear". In fact, "linear" here means linear in the *parameters* i.e., *in the beta's*. See the footnote on page 6 of my notes on M&M chapters 2 and 9 for more examples to make clear that the "linearity" is in the parameters and not in the X's.

" ... several *independent* variables on ... "

It would be better to say several "predictor" or "explanatory" or " stimulus" -- or better still -- "X" variables. The reason I emphasize this is that the term "independent" may be taken to mean that one can change one X without changing the other X. It is not always possible to do so. For example if we had X1 = year of birth, and X2 = age in the year 2000, it is not possible for X2 to vary independently of X1.

WHAT WE REALLY DID ON MARS

- **" because there were three discrete levels of water consumption"** (line 5)

The authors did not use the "discreteness " of C to fit the regression. They used it as a continuous variable just like height; they only exploited the discreteness to plot all of the data. We often use "slices" or "strata" to visualize the relation between one Y and 2 X's even if both X's are continuous: we choose for example three levels of X2 and the plot the relation between Y and X1 for each of these three levels of X2.

- **Figure 3-1 B**

Try to make this graph yourself with a spreadsheet such as Excel.

- **" the 3 lines can be thought of as contour lines... "** (2nd last para, p51 [55])

They think of "contours" differently! Usually, contours are used to show elevation (Y) as a function of say X1=longitude and X2=latitude in a topographical map: one connects (X1,X2) points which have the same value of Y. Here they're not doing that. See below. See also the use of colours in the example on the web page called "Average Weight as function of Height and Age" : the colors designate different values of Y, and the horizontal and vertical axes represent X1 and X2. The same idea is used in digitized x-rays or CT scans -- gray scales are used to show the intensity (Y) at each pixel.

Linking equations to planes is helpful. But the first priority should be to show equations which link the *mean* values of W to their corresponding values of height and water consumption, using some function of height and water consumption. In other words, the object of study is the relationship of the *conditional Y means* to the values of X1 and X2 that accompany these Y's. To be fair, the authors do say that figure 3.1 B shows an "alternative presentation" of the data. All I'm quibbling with is their use of $\hat{W}$ instead of "estimate of *mean* value of W at given values of X1 and X2".

- **"an equation which relates a dependent variable to several independent variables ... "** (text line 2, page 52 [line 1 p 56]).

Again, I would prefer that -- at least in the early chapters -- they continue to stress that they're looking for the relationship (systematic) between the *conditional means* of the dependent variable and several "independent" variables.

- **4 "population characteristics" or "assumptions for multiple linear progression"** (p 54 [58])

Some texts "go to town" on these assumptions, making them so forbidding that students are then afraid to even use regression. I have the same "take these with a grain of salt" comment here that I already made for simple linear regression.

If I were to re-emphasize my concerns about just one of these, it would be about the "**normality**". Beginning students -- and even some seasoned old-timers -- continue to misinterpret the so-called "requirement" of "normality". They plot the marginal (unconditional) Y's and look the see if they are Gaussian. They fuss if they are not. **What are assumed to be "normally" distributed are the ε's -- the individual variations about the specific *means at the different X levels.***

Suppose one aggregates all the Y's from the various X "slices" or "strata" or "X-locations" or "X-addresses". Then, even if the *conditional* variations are Gaussian, the distribution of the *aggregated* Y's will not be -- it is a mixture of different Gaussian distributions!

A simple example: In an ethnically homogeneous adult population:

1. The heights of *males* would be close to Gaussian.

2. The heights of *females* would be close to Gaussian.

3. The heights of *persons* would NOT be close to Gaussian.

4. BUT, the "requirement" of Normality is met!!! -- see 1. and 2. !!

- **How to fit the best plane through a set of data** (page 54 [58])

Page 54 deals with the *universe* of *possible* values of Y's at each X combination. It's only when we get to page 55 that we get into *estimating* the parameters that link these distributions together through the X's. Note the use of the *Greek* letters β_0 β_1 β_2 (*parameters* ..) on page 54 -- and the use of regular (Arabic) letters b_0 b_1 b_2 (*statistics* .. from samples) when we come to *data* on page 55.

- **Computing the regression coefficients** (page 55 [59])

Remember what it is we're trying to estimate: 3 beta's and 1 sigma. Unfortunately, many textbooks think the job of estimation is done once they have computed the b's.

We fit the "plane of means "to obtain estimates of β_0 β_1 β_2 and we calculate the "lack of fit" i.e. the SS_{res}, to obtain an estimate of σ. The b's and the $s_{y|x}$ have a similar form to those for simple linear regression in Chapter 2

- **" ... physical interpretation"** (last paragraph of page 55 [59])

- **" the *fact* that the three lines in figure 3.1A corresponding to the three levels of water consumption have the same slope illustrates the point that the *effects of changes in height on weight were the same at each level of water consumption*"**

Lines will only be parallel if the authors forced them to be parallel! Thus, it is not quite right to claim that just because the authors fitted three parallel lines, the effect of changes of height on weight was the *same* in each level of water consumption. That's an *assumption* one builds in as soon as a one chooses to fit three parallel lines.

- **"0.28 g per cm H is the *increase* in weight for each unit of *increase* in height holding water consumption constant"**

It is better to say that "0.28 is the *difference* in weight for each unit *difference* in height holding water consumption constant".

In addition, the 0.28 slope is the same 0.28 slope *irrespective of the level of water consumption*. This says that we have the "***same slope for different folks***". Later on, near the end of this chapter (page 94), we will encounter situations where the slope of μ_Y on X1 is not the same for all levels of X2, i.e., where the slope *does* depend on the level of the second of variable X2 --we have "***different slopes for different folks***".

The same applies to the 0.11 g per cup of water. This 0.11 is *assumed* to be constant over all levels of height. If one had sufficient data, one could see if the data support this assumption.

- **The variability about the regression plane**

The authors correctly emphasize the two essential components of regression, the *location* and the *spread* (variability). The locations are estimated by the fitted "plane of means", and the spread or variability is estimated by $s_{y|x}$.

- **Why divide by n - 3 here?**

For the same reason that we divide by n - 2 back in Chapter 2 and by n-1 back in course 607!

In 607, the focus was on estimating a single μ_Y and the variation about this μ_Y . The n residuals about the sample mean [the deviations from $\bar{y}$] had one (1) constraint: they had to add to zero.

When, in Chapter 2, we fit a line to estimate a "line of means", the n residuals (the deviations from the *line of means*) are now constrained in 2 ways -- so only n - 2 residuals are free to vary independently. Put another way, we have only *n - 2 independent assessments of variation*.

Now, in Chapter 3, we fit a plane. To do so we have to estimate 3 parameters. The n residuals are now constrained in 3 ways -- so only any n - 3 of them are free to vary independently. *We have three n - 3 independent assessments of variation.*

• Standard errors of the regression coefficient

"We assumed that the underlying population is normally distributed about the lane of means therefore all possible values of each of the b's will be distributed normally"

We could restate this by saying the ε's vary about the plane of $\mu_{y|x1\ x2}$ so the b's will cary around the β's with standard deviations that reflect (1) the amplitude of the ε's (2) the spread of the X's and (3) the sample size n.

The Gaussian-ness of the distribution of the b's can also be justified even if the ε's do not have a Gaussian distribution -- provided that the sample size is sufficiently large that the Central Limit Theorem comes into force.

• Expression for standard errors of the estimated slopes (b's) (equation 3.6)

Again, it would be more intuitive to rewrite equation 3.6 as

$$SE[b_i] = \frac{s_{y\ |x1\ x2}}{\sqrt{n} \times SD(X_i) \times \sqrt{1 - r^2_{\ x1\ x2}}} \qquad [\ approx:\ \sqrt{n} \approx \sqrt{(n\text{-}1)}\ ]$$

The standard error is now function of 4 factors, 3 of which we have seen before when we were dealing with a single X in Chapter 2.

The new factor is $\sqrt{1 - r^2_{\ x1\ x2}}$. To understand what it is and what its impact will be, let us first examine its structure. Since $r_{x1\ x2}$ is between -1 and 1, then $r^2_{x1\ x2}$ is between 0 and 1, so $\sqrt{1 - r^2_{\ x1\ x2}}$ is between 0 and 1; it is on the bottom of the expression.

Now, consider first the case where $r^2_{\ x1\ x2} = 0$ (or close to 0). This would apply if the distribution of X2 values is about the same for each value of X1 and vice versa [as would be the case in a clinical trial of the "treatments" X1 = 0/1, in which randomization was successful in making the distribution of X2 the same for the "X1 = 1" group as the "X1 = 0" group]. In this case, there is no alteration to the standard error.

But what if r = 0.6, say? Then $\sqrt{1 - r^2_{\ x1\ x2}} = \sqrt{1 - 0.36} = 0.8$. Thus, $SE[b_1]$ is increased or "inflated" by a factor of 1/0.8 = 1.25 = 25% over what it would be if X1 and X2 were uncorrelated.

If -- worse still -- r = 0.8 say, then the inflation is a $1/\sqrt{1 - 0.64} = 1/0.6$ or 1.7 This "inflation" is the price to pay for the fact that there is not enough separate variation in X1 and X2 to allow the true effect of X1 to be isolated well from the true effect of X2. (in the Belfast Catholic and Protestant story, r = 1, so the standard error of b_1 is infinite.

• Sample size for comparison of two means .. in presence of a "confounder" (up from 607)

Students learn in 607 how to calculate a sample size for a given power or power for a given sample size, when the interest is in a single difference of means $\mu_1 - \mu_0$, estimated by the "crude" $\bar{y}_1 - \bar{y}_0$. This crude difference in means can also estimated from the most fundamental of all simple regression situations. Let X=0 and X=1 denote the two groups being contrasted, so that the model is

$$\mu_{Y|X} = \mu_0 + \beta X$$

Then β represents the difference in means; so $b = \bar{y}_1 - \bar{y}_0$.

In 607, the following equation gives the minimum n's to "detect", with given "alpha" and "beta" error rates, a difference of $\Delta = \mu_1 - \mu_0$:

$$\text{number per group} = (Z_{alpha/2} + Z_{beta})^2 \times \frac{\sigma^2}{\Delta^2} .$$

What if we had a variable X2 which -- because of its imbalance with respect to the two treatment groups X1=1 and X1=0, and its own effects on Y -- we wished to include in the analysis, using the model

$$\mu_{Y\,|\,X1\,X2} = \mu_0 + \beta_1 X_1 + \beta_2 X_2 .$$

From equation 3.6, we can now **generalize** the sample size formula

$$\text{number per group} = (Z_{alpha/2} + Z_{beta})^2 \times \frac{\sigma^2}{\Delta^2} \times \frac{1}{1 - r^2_{x1\ x2}}$$

You can think of the $\frac{1}{1 - r^2_{x1\ x2}}$ as the "**variance inflation factor**" or equivalently -- since sample size and variance are "exchangeable" -- as the "**sample size inflation factor**" to compensate for the correlation of X1 and X2 (i.e., the imbalance of X2 with respect to X1).

• Some additional links involving SE's

In the "607-level" example, the interest is in a crude estimate of difference of means $\mu_1 - \mu_0$. This difference in means can be estimated the most fundamental of all simple regression situations. Let X=0 and X=1 denote the two groups being contrasted, so that the model is

$$\mu_{Y|X} = \mu_0 + \beta X$$

with β representing the difference in means.

In 607, this difference was estimated

	"directly"	"by regression"
as	$\bar{y}_1 - \bar{y}_0$	b
and its SE is calculated as	$SE[\bar{y}_1 - \bar{y}_0] = s\sqrt{1/n_1 + 1/n_0}$	$SE[b] = \dfrac{s}{\sqrt{n} \times SD[X]}$
where s is	SD of the pooled within-group deviations	.SD of "residuals"

But, using π_1 and $\pi_0 = 1 - \pi_1$ to denote what fractions of the entire $n = n_1 + n_0$ subjects are in group X=1 and X=0 respectively, we can re-write $SE[\bar{y}_1 - \bar{y}_0]$ as

$$SE[\bar{y}_1 - \bar{y}_0]$$

$$= s\sqrt{1/(n\,\pi_1) + 1/(n\,\pi_0)}$$

$$= s\sqrt{1/n \times 1/(\pi_1\pi_0)}$$

$$= \frac{s}{\sqrt{n} \times \sqrt{\pi_1\pi_0}} = \frac{s}{\sqrt{n} \times SD(X)} = SE[b]$$

Muddying the water: multicollinearity

The effects of the " instability " of the fitted regression coefficients are summarized verbally in the first paragraph. Here is a table of the effect on the standard error (SE) using as a "base" the SE when r = 0. (The degree to which the sample size would need to be increased to counteract this instability is a function of *variance*, not SE, and so the tabulated SE inflation factors would need to be *squared*).

The "SE inflation factor" $1/\sqrt{1 - r^2_{x1\ x2}}$
tabulated as a function of r

$r_{X1,X2}$	$1/\sqrt{1 - r^2_{x1\ x2}}$	$r_{X1,X2}$	$1/\sqrt{1 - r^2_{x1\ x2}}$
0.0	1.00 (base)	0.7	1.40
0.1	1.01	0.8	1.67
0.2	1.02	0.85	1.90
0.3	1.05	0.90	2.29
0.4	1.09	0.95	3.20
0.5	1.15	0.98	5.02
0.6	1.25	0.99	7.09

- **"This situation, called multicollinearity is one of the potential pitfalls in conducting a multiple regression analysis"** (last sentence of first paragraph)

Technically speaking, multicollinearity is when one variable, say X1 is predictable from some linear combination of the remaining X's. Here, there is only one other X, so we're simply dealing with the correlation of X1 and X2. If there are multiple X's, then it is the correlation of X1 with some *combination* of the others that is called multicollinearity.

Note that whereas correlation of X1 with X2 is easy to see in a scatter plot, the (multiple) correlation of X1 with a combination of the other X's is not so easy to visualize.

Use the interactive Excel program (on web page) to see the effects of the correlation between X1 and X2 on the b_1 and b_2 estimates.

- **Figure 3-3 legend**

"This situation, called multicollinearity, occurs when the two independent variables contain the same information"

Again here, because there just 2 variables, it would not be enough to say that we have **collinearity** between X1 and X2. *Multi*collinearity is between one X variable and some or all of the other X's.

" There is not a broad 'base' of values of the independent variables X1 and X2 to support the regression plane"

Try balancing a rigid plane or sheet of paper on a long narrow bed of inverted nails. The slightest shift in one of the nails will create instability and will rock the plane. I've called the Excel program "hammock" because it is as though the hammock is being supported by a very narrow base rather than from all four corners. Some texts referred to this instability as balancing a plane on a "*knife-edge*".

Multicollinearity: is it all bad?

Not if the objective is *prediction*. *Yes* if one wishes to *separate or "partial out" the effects of various X's one from one another*.

Figure 3-4

Again here the authors seem to imply that multicollinearity involves one X with just one other X. In fact, as I emphasized above, it can be one X with *a combination of several other* X's.

"data falling in a cigarette shaped cloud"

The cigarette data on the web page are a useful "back here on earth" data set that show the same pattern as in Figure 3-4. Two characteristics of each cigarette brand, X1=the amount of tar and X2=the amount of nicotine, predict the amount of carbon monoxide (Y) produced by the cigarette. Plot the data in 3-D for yourself.

DOES THE REGRESSION EQUATION DESCRIBE THE DATA?

"... testing the significance of the regression coefficient" (second line)

Even in the case of a single independent variable, there is a big difference between the "slope b being statistically significant" and "the equation describing the data *well*". In simple linear regression, the tests of significance performed by default by packages are in relation to the null hypothesis H_0: $\beta = 0$. Even if the test is "positive" (i.e., b is significantly different from 0), β could still be minuscule: b could simply be "statistically" significant because the sample size is large. Even if β is truly sizable, there can still be substantial unexplained variation: the equation may fit the *means* reasonably well, but there may still be a lot of individual variation around the specific means.

For example, it might be that there is a perfectly linear relationship between the *means* of Y = blood pressure and X = age, but individuals would still vary considerably about the age-specific means.

Even worse is the (naive) conclusion -- from the "significant F test" (p = 0.02!!!) -- that the simple linear equation 2.21 on page 43 fits the Gray Seal data "well".

The same caveats apply to tests in a regression with 2 X terms. A "significant" regression does not necessarily fit the individual datapoints "well"; all one can conclude is that the variables in the equation are "better than nothing", i.e. that

$$\mu_{Y \mid X1\ X2} = \mu_0 + \beta_1 X_1 + \beta_2 X_3$$

is better than

$$\mu_{Y \mid X1\ X2} = \mu_0$$

A better heading for this section might have been "tests of a (*composite*) null hypothesis" i.e. concerning *several* β's. The test at the bottom of page 64 is in relation to the null hypothesis

$$H_0:\ \beta_1 = 0 \textbf{ and } \beta_2 = 0$$

Whereas the test procedure is described in detail on page 64, what is missing is an explicit statement of the null hypothesis being tested, and what the alternative hypothesis is. In fact, the alternative hypothesis is that at least one of the two β's is non-zero:

$$H_{alt}:\ \beta_1 \neq 0 \textbf{ or } \beta_2 \neq 0 \text{ (includes possibility that } \textbf{both } \beta_1 \neq 0 \text{ and } \beta_2 \neq 0)$$

The textbook by KKMN makes a point of carefully stating all hypotheses.

" to test the overall goodness of fit of the regression plane given by equations 3.1 and " (line 4, page 65 [last para, p 68])

A more correct way to say this would be to test "whether one or both of the variables help predict..."

Some authors (as do the authors themselves later) call this an ***overall*** test

"Incremental sums of squares and the order of entry" (p 65 [69])

This is called the "extra some of squares" in some textbooks

Some software packages, such as SAS, report different tables under the headings of "Type I sums of squares" and Type III sums of squares". Below and right, for example, are the outputs from INSIGHT.

Analysis of Variance

Source	DF	Sum of Sq.	Mean Sq.	F Stat	Prob > F
Model	2	54.213	27.107	1591.8	0.0001
Error	9	0.153	0.017		
C Total	11	**54.367**			

Type I Tests

Source	DF	Sum of Sq.	Mean Sq.	F Stat	Prob > F
H_HEIGHT	1	47.839	47.839	2809.2	0.0001
C_WATER	1	6.375	6.375	374.3	0.0001

Type III Tests

Source	DF	Sum of Sq.	Mean Sq.	F Stat	Prob > F
H_HEIGHT	1	16.368	16.368	961.2	0.0001
C_WATER	1	6.375	6.375	374.3	0.0001

One way to tell a table of Type I sums of squares (effects of variables added *in order*) from a table of Type III sums of squares (effect of variables if *added last*) is if the sums of squares for the independent variables in the table add up to the "regression" or "model" sum of squares . If they do, then one is dealing with Type I or "sequential" sums of squares.

In the Type I table, 47.8... + 6.3.. = the **54.2** in the overall Table; in the Type III table, 16.3... + 6.3.. ≠ 54.213 in the overall table.

If, later on, you forget which is Type I and which is Type III (who could blame you), you can still remember that if they add up, they are sequential! The term "SEQ SS" (for "sequential" SS) is decidedly more descriptive than "Type I" SS.

Some texts use the following notation for sequential SS.

SSreg = SS(H) + SS(C | H)

In Figure 3.2, the sequence for the incremental sums of squares (Type I SS in SAS) is "H then C"

SSreg = SS(X2) + SS(X1 | X2)

From SS(H) & SS(C | H) , can we calculate SS(C) + SS(H | C)?

Q: What if we wish to reconstruct the sequential sums of squares for C, then the incremental effect (SS) of H given C, but forgot to put the variables into the regression in that particular order? Can we get there from the printout in figure 3.2?

A: No! One would need to rerun the program with the variables listed in the correct order, i.e., C entered before H

Type I Tests

Source	DF	Sum of Sq.	Mean Sq.	F Stat	Prob > F
C_WATER	1	37.845	37.845	2222.3	0.0001
H_HEIGHT	1	16.368	16.368	961.2	0.0001

Relationship to t tests of individ. regression coefficients (p 67/71)

" ...t test on a particular X is equivalent to conducting an F-test based on the incremental sum of squares with that particular variable put into the equation last"

Check this in the Type III SS table ("variable added last") above

C: T Stat = 19.3.. $(\text{T Stat})^2 = 19.3..^2$ = 374.34 = F Stat (Type III)
H: T Stat = 31.0.. $(\text{T Stat})^2 = 31.0..^2$ = 961.20 = F Stat (Type III)

Note that if one is interested in the contributions of variables if they were added last, it doesn't matter which order one puts them in the equation.

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob > |T|
INTERCEPT	1	-1.220	0.3210	-3.8	0.0042
C_WATER	**1**	**0.111**	**0.0057**	**19.3**	**0.0001**
H_HEIGHT	**1**	**0.283**	**0.0091**	**31.0**	**0.0001**

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob > |T|
INTERCEPT	1	-1.220	0.3210	-3.8	0.0042
H_HEIGHT	**1**	**0.283**	**0.0091**	**31.0**	**0.0001**
C_WATER	**1**	**0.111**	**0.0057**	**19.3**	**0.0001**

"As already discussed, the standard errors (SE's) for the b's take the correlation of the X's into account..."

Recall the standard error of b_i using the formula

$$SE[b_i] = \frac{S_{y\,|x1\ x2}}{\sqrt{n} \times SD(X_i) \times \sqrt{1 - r^2_{x1\ x2}}}$$

The correlation between the two X's appears in the denominator.

Tested by the "T Stat" b_1 / $SE(b_1)$ is the model

$$\mu_{Y\,|\,X1\,X2} = \mu_0 + \beta_1 X_1 + \beta_2 X_2$$

versus

$$\mu_{Y\,|\,X1\,X2} = \mu_0 + \mathbf{0} \times X_1 + \beta_2 X_2$$

and by the "T Stat" b_2 / $SE(b_2)$ is the model

$$\mu_{Y\,|\,X1\,X2} = \mu_0 + \beta_1 X_1 + \beta_2 X_2$$

versus

$$\mu_{Y\,|\,X1\,X2} = \mu_0 + \beta_1 X_1 + \mathbf{0} \times X_2$$

" ... b / SE(b) tests whether X contains significant predictive information about Y after taking into account all the information in the other independent variables" (page 68 [72])

This is the same as if this particular X were entered last. Note: T^2= F here. The T statistic is more informative since it gives the sign of b.

THE COEFFICIENT OF DETERMINATION AND THE MULTIPLE CORRELATION COEFFICIENT

The authors don't make much of the fact that the multiple correlation coefficient is the correlation coefficient of the Y's with the fitted Y's. In fact, one could turn this around backwards and ask: what if one set out to find the linear combination of X's which gave the maximum correlation with the Y's. In fact, the coefficients in this linear combination would turn out to be the least squares estimates (the b's).

" it is possible to construct hypothesis tests for overall goodness of fit based on the value of R^2 ..."

I hesitate to sell them as tests for the "overall goodness of fit"

In the social sciences, most testing (of effects when variables are added in order and added last) is carried out on the increments in R^2 rather than on the b's. In the biomedical field, we usually more interested in the β's because they give us some sense of magnitude and often have a physical interpretation.

MORE DUMMIES ON MARS

They phrase their question : "*does exposure to secondhand tobacco smoke affect the relationship between height and weight of Martians?*"

While this is a legitimate question, it is **not the question they addressed in their analysis**. The method for answering the question in italics above belongs in the section on "interactions" starting on page 94.

The question they address in this section is -- as they themselves note in the footnote to page 72 -- "*is there an effect of secondhand tobacco smoke on weight, controlling for the effect of height?*" Or, put slightly differently, "*does exposure to secondhand smoke affect the weight of Martians, all other factors (here height) being equal?*"

Is Glantz's use of "dummies" and "smoking" an intended *double entendre*?

"the weight-height curve is shifted down by a constant amount for the "dummies", no matter what height there are" (page 70 [74])

This is a consequence of the form of the regression equation adopted at the bottom of page 69; i.e., the assumption of a parallel shift is "built in" by the authors.

Note that it would be good if the data bore out this assumption, since it makes it much easier to report the results for smoking -- one does not have to give a separate estimate of the smoking effect for Martians of different heights. This allows it to be a simple "one effect fits all" story.

Note the primary interest on the effect of secondhand smoke. The fact that individuals are of different heights is more of a nuisance.

Obviously, with height being such a strong determinant of weight, one wishes to make sure the comparison of smokers and non-smokers is fair ("balanced", a comparison of "like with like") with respect to height. But, beyond that one is not interested in height *per se.*

" we conclude that exposure to secondhand smoke stunts Martians' growth" (end of the first paragraph page 72 [75])

It is clear from this conclusion that the authors did not intend the question the way they posed it on page 69.

See the "making comparisons fairer" section of the article on "Appropriate uses of multivariate analysis" referred to in the notes on Chapter 1.

See also the chapters by Anderson, Anderson *et al.* accessible via the web page.

Few textbooks emphasize this "analysis of covariance" sufficiently. KKMN put it quite late in their text. More often than not, the focus of multiple regression methods in the biomedical sciences is the effect of one specific variable while adjusting for other (confounding) variables. So why not make more of this early on?

GENERAL MULTIPLE LINEAR REGRESSION

The main changes in going from two X's to k X's are:

- The calculations become more involved. It *is* possible to carry out multiple linear regression using a program or calculator which performs simple linear regression, but it takes programming, organization and patience! If you want a taste, see my notes on "multiple regression as a series of simple regressions" in the "notes on multiple linear regression from 607". The example involves just 2 X's: see if you can generalize it to 3 or more X variables.

- We cannot plot the data in all of the dimensions.

- To get an unbiased estimate of the variation of the variation ($\sigma^2_{Y \mid X1,\ldots Xk}$) one divides the SS_{res} by $n - (k + 1)$. The reason for this divisor is that one has fitted (k + 1) b's to n data points

- $SE[b_i]$ now involves in its denominator the multiple correlation of X_i with the best linear combination of the (k-1) other X's.

- The "overall F test" now involves the null

  H_0: $\beta_1 = 0$ **and** $\beta_2 = 0$ ... **and** ... $\beta_k = 0$

  vs. the alternative hypothesis that at least one of them is non-zero:

  H_{alt}: $\beta_1 \neq 0$ **or** $\beta_2 \neq 0$ ... **... or ...** $\beta_k \neq 0$

  (includes possibility that **all** β's $\neq 0$)

Again, as discussed above, these are seldom interesting hypotheses.

When one has k variables, the "partial F test" of the null

H_0: a **subset** of 1 (or more) of the k β's is (are) 0

[say subset is of size "s"]

against the

H_{alt}: at least 1 of the β's in the subset is non-zero:

involves the (partial) F statistic

$$F = \frac{MS_{EXTRA}}{MS_{residual}} = \frac{\text{Difference in } SS_{REG} \,/\, s}{SS_{residual} \,/\, [n - (k+1)]}$$

and comparing it against the $F_{s,n - (k+1)}$ distribution.

The Overall F test is at one extreme ("subset" is *all* k X's) and the t-test (or its square, the "F with 1 df" test, for the variable added last) is at the other (subset involves just 1 X).

Chapter 8 of KKMN goes into these Partial F tests in some detail. I summarized them in session 4 in 1999. They use the notation "full model" and "reduced model".