Interpretation of regression coefficients: For the simple linear regression we know that the units of the intercept are the same as the units on the variable "Y". The slope has units of "Y units per X unit". The same is true of the intercept and slopes in multiple regression, except each X_{ij} could have different units so each estimated regression coefficient (b_i) would have different "Y units per X_i unit".

The regression coefficients in multiple regression have similar interpretations and limitations to those in simple linear regression. However, as mentioned, the individual variables X_i modify each other. As a result the interpretation depends on other variables included in the model.

Also, all of the assumptions of simple linear regression also apply to multiple regression. Normality and constant variance are assumptions that can be examined from the residuals. The additional assumptions of independence and "X_i measured without error" are harder to examine.

The text discusses what it calls "specifically constructed explanatory variables". These include polynomial models, which we have discussed, and indicator variables. Indicator variables are usually coded as the values "0" and "1". For a single binary indicator variable, only a single column is needed, although two could be used (see "Timing Indicators" below). Note that a single variable would have only 1 d.f. Although two columns can be used, this variable can only use a single degree of freedom.

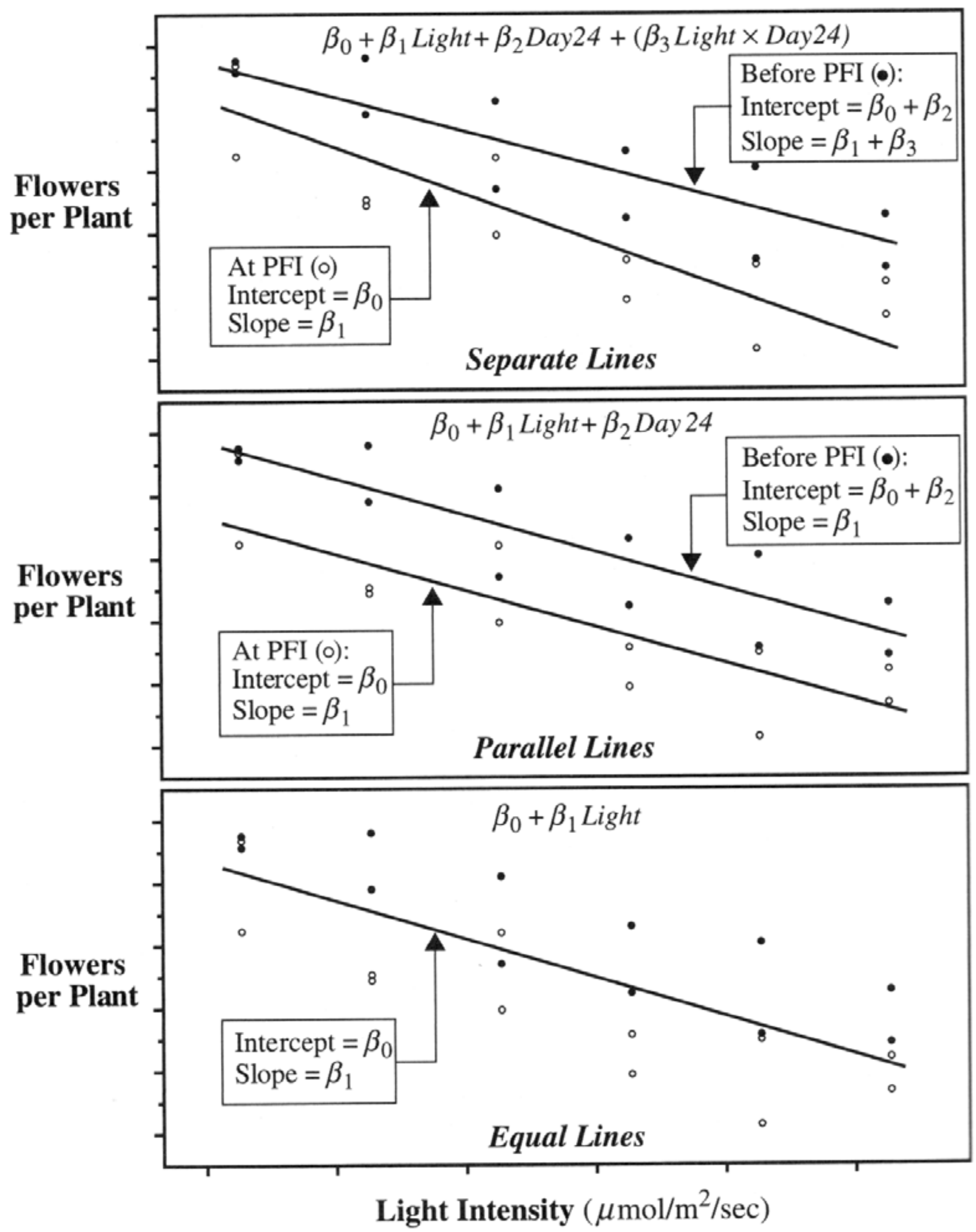
After Display 9.7 from the text. Explanatory variables coded as indicator variables and interaction.

Original variable			Timing indicators		Light level indicators						Interaction
FLOWERS	INTENSITY	TIME	day 24	day 0	L150	L300	L450	L600	L750	L900	Light x day24
62.3	150	0	0	1	1	0	0	0	0	0	0
77.4	150	0	0	1	1	0	0	0	0	0	0
77.8	150	24	1	0	1	0	0	0	0	0	150
75.6	150	24	1	0	1	0	0	0	0	0	150
55.3	300	0	0	1	0	1	0	0	0	0	0
54.2	300	0	0	1	0	1	0	0	0	0	0
69.1	300	24	1	0	0	1	0	0	0	0	300
78.0	300	24	1	0	0	1	0	0	0	0	300
49.6	450	0	0	1	0	0	1	0	0	0	0
61.9	450	0	0	1	0	0	1	0	0	0	0
57.0	450	24	1	0	0	0	1	0	0	0	450
71.1	450	24	1	0	0	0	1	0	0	0	450
39.4	600	0	0	1	0	0	0	1	0	0	0
45.7	600	0	0	1	0	0	0	1	0	0	0
62.9	600	24	1	0	0	0	0	1	0	0	600
52.2	600	24	1	0	0	0	0	1	0	0	600
31.3	750	0	0	1	0	0	0	0	1	0	0
44.9	750	0	0	1	0	0	0	0	1	0	0
60.3	750	24	1	0	0	0	0	0	1	0	750
45.6	750	24	1	0	0	0	0	0	1	0	750
36.8	900	0	0	1	0	0	0	0	0	1	0
41.9	900	0	0	1	0	0	0	0	0	1	0
52.6	900	24	1	0	0	0	0	0	0	1	900
44.4	900	24	1	0	0	0	0	0	0	1	900

For variables with more than a single dichotomous value, sets of indicator variables are used. The levels of the variable are still coded as "0" and "1" indicator variables, and one can be set up for each level of the variable. However, for t levels of a treatment only t – 1 degrees of freedom can be assigned. The default pattern in SAS is to omit the variable in the last alpha-numeric position. The table above shows that "light level indicators" could be coded as an indicator variable (rather than a quantitative variable).

The Meadowfoam flower experiment was presented in this chapter as a regression. I would normally treat this analysis as a "Two-way Analysis of Variance" or "2 x 6 Factorial treatment arrangement in a CRD (Completely Randomized Design)". I would do this because this analysis is a designed experiment with two treatments with a total of 12 combinations. Once fit as an ANOVA, the model could be tested for a linear trend in the quantitative treatment "light level" and eventually reduced to a regression with one or two slopes and one or two intercepts. This would be the analysis of covariance presented by the text.

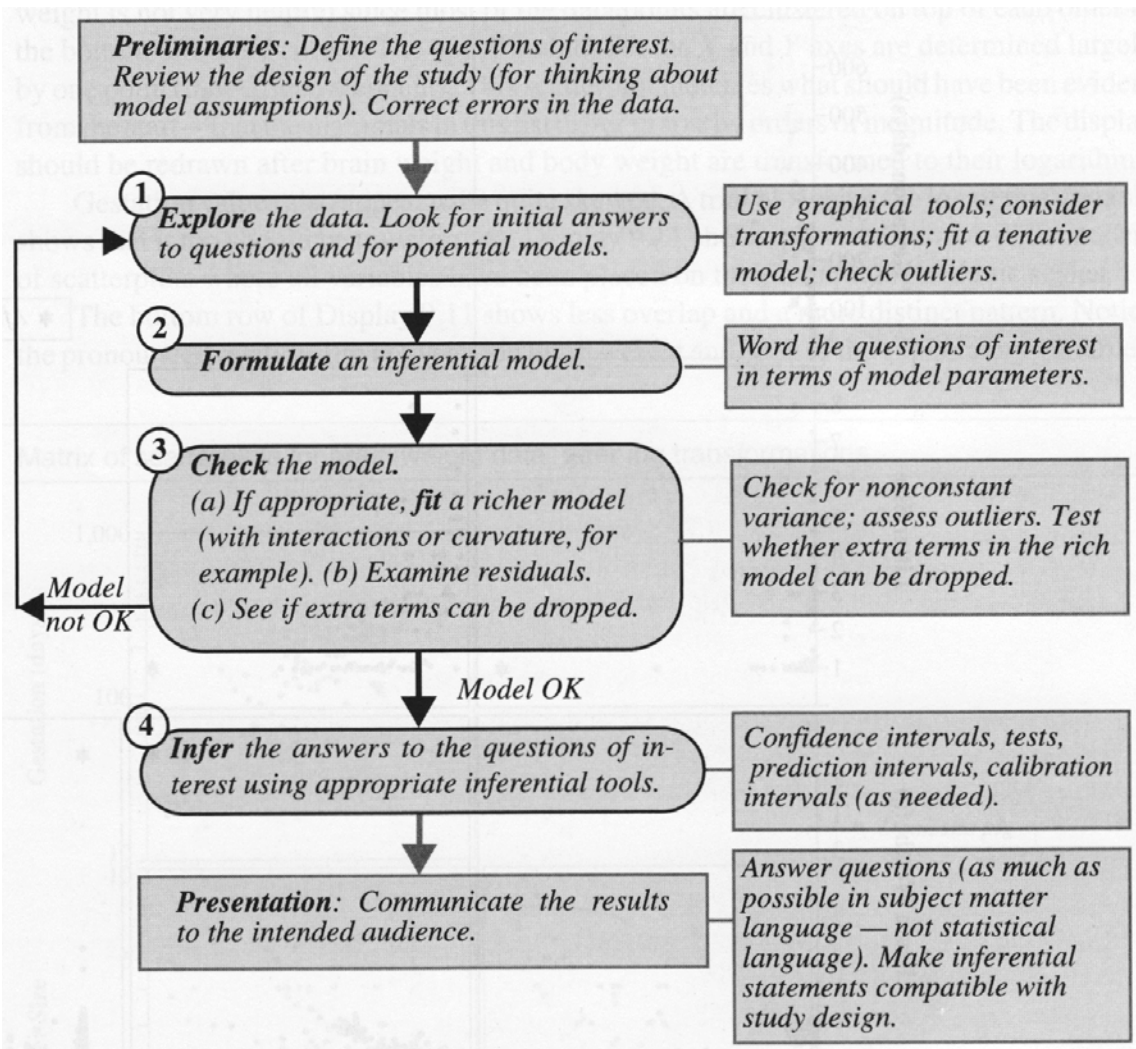
You have seen my graphical description of Analysis of Covariance (AnCova). Graphics from the text are given below. Note particularly the effect of the interaction between the variables. This interaction term was included in the table above.



The book suggests a general approach to fitting a model. I generally agree with the model for regression analysis. However, designed experiments generally have a predefined design, and I would start with that model. This is generally a "full" model with all components included. I would then examine potential reductions in the full model, like fitting a quantitative variable to light intensity instead of an indicator (or class) variable.

If we fit a model with timing (indicator variable) and light (as an indicator variable) and then fit the same model with light as a quantitative variable, what do we call the extra SS for this model?

For regression model I would be in general agreement with the book, starting with a simple model and then fitting a "richer" model where appropriate. I particularly like the ancillary comments: using graphics to check for transformations and outliers, stating your research question in terms of parameter estimates, checking assumptions and the significance of variables in the model, the use of confidence intervals (especially if you have stated your research question in terms of parameter estimates) and making your final discussion in terms of the subject matter and not the statistics.



The books refers to a "randomized experiment" as one where the investigator controls all of the conditions of the experiment, including the treatment and which individuals are assigned to which treatments. This is a designed experiment done under controlled conditions. As the book points out, in such an experiment, such as the light experiment and timing manipulations, the investigators can attribute the observed effects to the experimental manipulations because they control all other aspects of the experiment. If an investigator wishes to demonstrate "cause and effect", this is the type of experiment that must be done. The statistical examination of observational or survey type of data, such as the mammal brain size data, can yield some compelling correlations, but causal interpretations are not justified from this type of data.

When to include interaction – The book states that interactions are not normally included in regressions. I disagree in the case of analysis of covariance, I usually start with the interaction and work up the model from the bottom, reducing the model where possible. It is true that in most multiple regressions we do not automatically fit interactions to the variables.

I agree that if interactions to variables are included in the model the individual terms should be included in the model even if they are not significant. For example if the interaction of X_1 and X_2 is significant (X_1X_2) then X_1 and X_2 would both be included in the model even if their coefficients were not significantly different from zero.

The next example is not a designed experiment, it is an observational study. In this study biologists are interested in the correlations between brain size and certain variables. Those variables are gestation time, litter size and body size. Since I expect a power model would be best suited to fit brain size to body size I included logarithm transformed variables in the mix.