PREDICTORS, TERMS, AND CATEGORICAL PREDICTOR VARIABLES
IN MULTIPLE LINEAR REGRESSION

Multiple linear regression allows more possibilities than simple regression.

I) Terms that are functions of predictors.

Examples:

$E(Y|x) = \eta_0 + \eta_1 x + \eta_2 x^2$ (one predictor, 2 non-constant terms)

$E(Y|x_1, x_2) = \eta_0 + \eta_1 x_1^2 + \eta_2 x_1 x_2 + \eta_3 x_2^2$ (two predictors, 3 non-constant terms)

"*Linear regression*" means "linear in the terms" or "linear in the coefficients"

II) Categorical predictors (sometimes called *factors*)

Example: If sex (with two categories, male or female) is relevant in the context, we can introduce a *coding* (or *dummy* or *indicator*) variable:

$$X = \begin{cases} 0 \textit{ if female} \\ 1 \textit{ if male} \end{cases}$$

For example, if we are regressing W = weight on both sex and H = height, we might consider modeling the mean function as

$$E(W|H, \text{sex}) = \eta_0 + \eta_1 X + \eta_2 H + \eta_3 XH$$

This says:

$E(W|H,\text{female}) =$

$E(W|H, \text{male}) =$

The term XH is called an *interaction term*. (Why?)

We might also use coding variable $X = \begin{cases} -1 \textit{ if female} \\ 1 \textit{ if male} \end{cases}$

The choice of coding variable may depend on context or custom; the interpretation of coefficients will depend on the choice. (Try the above example with this new choice of X!)

→ *Read pp. 146 – 152 for a variety of examples of using terms and coding variables.*

Note: A categorical variable with more than two categories can be coded using more than one coding variable; more on this later.