**9.0 – Terms vs. Predictors**In multiple regression we study the conditional distribution of a response variable () given a set of potential predictor variables . As dicsussed previously, these variables can have any data type – continuous/discrete, ordinal, or nominal. In this section we focus on the concept of terms. Terms in a multiple regression model are functions of the predictors . We will denote the terms in a multiple regression model using to avoid confusion with the predictors ().

**9.1 – Multiple Linear Regression Model**The form of the multiple regression model is given by

and typically we assume .

In matrix notation,

where,

 in our regression model and the are the observed values of the *jth* term. As before, the OLS estimates of the parameters are found using matrices as:

Thus specifying the mean function part of the model involves deciding what terms to include in the model! For example, in **Example 8.1** we considered the several models for the selling price () as a function of .

 Note: These are **terms** in these models.

**9.2 - Types of Terms**

As we saw in Example 8.1 terms can be the predictors themselves, as in the case of , or a function of a predictor as in the case of Recall, and if . Below we give the most commonly used terms in a multiple regression models. These are the basic “*building blocks”* of a multiple regression model.

**Intercept term**

 🡨 This term gives the column of 1’s in the model matrix . We do not
 need to include an intercept term, but we almost always do!

**Predictor terms**

 🡨 Terms can be the ***predictors*** themselves as long as the predictor is
 meaningfully numeric! (i.e. count or a measurement).

**Polynomial terms**

 🡨 These terms are integer powers of numeric predictors, e.g.

**Transformation terms**

 🡨 Here is the Tukey Power Transformation Family.

**Interaction terms**

 🡨 Here the term is a product of two terms (, where these
 two terms could be of any other term type.

**Dummy terms**

Other examples of **:**

These terms generally come from a dichotomous
nominal predictor as in the case of Fireplace? (Y/N)
in Example 8.1.

**Factor terms (nominal or ordinal variables with more than 2 levels)**

Suppose the predictor is a nominal or ordinal variable with levels (. Then we chose one of the levels as the ***reference group***and create dummy terms for the remaining
 levels.

E.g. Suppose

We could choose *4 = Oil* as the reference group and create dummy variables for the other two manufacturers, i.e.

Why wouldn’t we create a dummy variable for each level of the levels of the variable? For this example that would mean also creating,

The problem with doing this is that the sum of the columns corresponding to these terms would be a column of ones. When one column in the matrix is linear combination of other columns in the matrix then the inverse of the matrix doesn’t exist, i.e. the matrix is ***singular***.

**Example 9.1 – Saratoga, NY Homes and Fuel Type (Datafile: Saratoga, NY Homes.JMP)**

As an example consider the regression of selling price on fuel type which is a nominal variable coded as (2 = Gas, 3 = Electric, 4 = Oil). Below is a portion of the data table from JMP with these variables. I have created three dummy variables, one for each fuel type.



If we perform the regression of selling price (Y) on all three of the dummy variables, one for each fuel type this is what happens in JMP:



If we simply put Fuel Type as coded (2 = Gas, 3 = Electric, 4 = Oil) into the model, JMP will automatically drop one of the levels, 4 = Oil in case, and estimate parameters associated with the dummy variables for the other two fuel types.





 **Note:** Regression on a single nominal or ordinal variable with levels is equivalent to one-way ANOVA covered in STAT 310 and STAT 365. Furthermore, one can show ANOVA (of any kind) is really just regression on factor terms.

**Important Note on Coding for Nominal and Ordinal Predictors**

It is important to realize that when working with nominal or ordinal predictors in JMP that the default coding is (-1,+1), i.e. contrast coding, NOT (0,1), i.e. dummy or indicator function coding. However, you can select the **Response *Name* > Estimates > Indicator Parameterization Estimates** option to obtain parameter estimates using dummy coding which I highly recommend doing.



R uses dummy variable coding as the default! Thus when developing multiple regression models I generally use R for this and other reasons. I will be giving you a very thorough handout/tutorial on performing multiple regression in R shortly.

**9.3 – Summary of Multiple Linear Regression (MLR) – predictors and terms**

As stated above the general form of the multiple regression model is given by

where the are the model terms created from the predictors . We also typically assume that the .

To develop a multiple linear regression (MLR) model we need to determine what terms to create and include in our model for the conditional mean, . This may seem like a daunting process as the possibilities are seemingly infinite, especially when we have a large number of predictors (i.e. is large), however there are a number general guidelines and tools that we can use help us in the model development process. In subsequent sections will look focus more closely on some of the term types discussed in this section. In the next section we will focus on cases where the predictors are primarily numeric (continuous/discrete), in Section 11 we will focus on factor and interaction terms, and in Sections 14 & 15 we will focus on response and predictor transformations respectively.