1 Multiple Regression Analysis

New challenges: 1) More difficult to choose a best model, 2) More difficult to visualize the model, 3) Sometimes more difficult to interpret, 4) Extra computing is required. Written as

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon \text{ or as } y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \varepsilon_i.$$

Some independent variables may be higher-order terms, such as $x_{i2} = x_{i1}^2$. It is still easy to visualize multiple regression models when all terms are higher-order terms of a single variable. Otherwise the model represents a surface or hyper-surface.

2 Assumptions

Linearity, Independence, Normality, and Homoscedasticity, similar to those from simple linear regression. Note that the assumptions apply to $y \mid x_1, x_2, ..., x_k$, and not to y. We can use $\widehat{\varepsilon}_i = y_i - \widehat{y}_i$ values to assess model assumptions. To estimate the β values, we again use the estimates that minimize the sum of squared errors (SSE). These can be written in matrix form:

$$\widehat{\beta} = (X\,'X)^{-1} X\,'Y.$$

The ANOVA table is interpreted as before, including the use of

$$R^2_{y\,\cdot\,x_1, x_2, ..., x_k} = \frac{SS(\text{Total}) - SS(\text{Residual})}{SS(\text{Total})}.$$

3 Regression Terminology

The **degree** of a term is the sum of exponents for its x variables. Thus the degree of x_i^p is p, and the degree of $x_i^p x_j^q$ is p+q. The **order** of a regression model describes the degrees of its terms. A **first-order model** has only terms of degree 1, while a **second-order model** has all possible first degree and second degree terms. Interpretation of slopes: recall that for the simple linear regression model $y = \beta_0 + \beta_1 x_1 + \varepsilon$ the interpretation of β_1 is that it describes the amount of change in $E(y) = \mu_{y|x}$ given a unit change in x. For the first-order multiple regression model $y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon$, we have a **partial slope** interpretation: that β_1 describes amount of change in $E(y) = \mu_{y\,\cdot\,x_1, x_2, ..., x_k}$ given a unit change in x_1, assuming that all other x_i terms are held constant. This interpretation is not always possible when second-degree terms are in the model. A **dummy variable** is a variable that only takes the values 0 and 1. For example, we could define an x variable to be equal to 1 for hot sandwiches and equal to 0 for cold sandwiches in our sandwich example. We will see later that there is a connection between dummy variables and **ANOVA**.