Handout 03: Matrix Computations and Multivariate Regression

Linear models are amongst the most well known and often-used methods for modeling data. They are employed to study the outcomes of patients in clinical trials, the price of financial instruments, the lifetimes of fruit flies, and many other responses from a wide range of fields. Their popularity is not unwarranted. In fact, the discussion of linear models and their variants take up a considerable portion of this text.

Consider observing n pairs of data (x_i, y_i) for $i = 1, \dots n$. A simple linear model would assume that the data are generated according to the equation

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{1.1}$$

where ϵ_i is some unobserved error term and the β_j's are unknown constants. The goal of statistical modeling is to use the observed data to, in some fashion, estimate the parameters β_0 and β_1.

Why are linear models so popular? One important attribute is that linear models provide a concrete interpretation for all of their parameters. Take the two variable model for predicting housing sale prices as a function of total area (in square feet or square meters) and the number of bedrooms,

$$\text{price}_i = \beta_0 + \beta_1 \cdot \text{area}_i + \beta_2 \cdot \text{bedrooms}_i + \epsilon_i. \tag{1.2}$$

The parameters in this model tell us how much the response, price, changes when one of the predictor variables changes with the other variable held fixed. Mathematically, we can describe this precisely using partial derivatives

$$\beta_1 = \frac{\partial \, \text{price}}{\partial \, \text{area}}, \tag{1.3}$$

$$\beta_2 = \frac{\partial \, \text{price}}{\partial \, \text{bedrooms}}. \tag{1.4}$$

The model separates the effect of the total size of a house and the total number of bedrooms. This information is useful to real estate agents, homeowners, construction companies, and economists. Linear models also allow for the interpretation of categorical predictors through the use of indicator variables. If our housing price data also includes information about whether a given observation is from one of three neighborhoods, say 'uptown,' 'downtown,' and 'suburbia,' we can define variables that are one when observation i is in the given neighborhood and zero otherwise. A linear model with these variables may be written as

$$\begin{aligned} \text{price}_i = \beta_0 + \beta_1 \cdot \text{area}_i + \beta_2 \cdot \text{bedrooms}_i + \\ \beta_3 \cdot \text{downtown}_i + \beta_4 \cdot \text{uptown}_i + \epsilon_i. \end{aligned} \tag{1.5}$$

The parameter β_3 can still be viewed as a partial derivative, here representing the difference in the expected price between a house in suburbia and a house in the downtown neighborhood, if both are the same size and have the same number of bedrooms.

The relatively simple form of linear models allows for a great deal of variation in the model assumptions. The x_i's can be treated as fixed values, a *fixed design*, or they may be considered to be random variables themselves, as in a *random design* model. In biological applications the analysis usually depends on strict independence between the errors. In time series data, as commonly seen in finance or macroeconomics, the ϵ_i are often serially correlated with one another. Linear models such as the autoregressive integrated moving average (ARIMA) model and the autoregressive conditional heteroskedasticity (ARCH) model are used to model time series data with serial correlation structures. Longitudinal medical studies, where data is collected on multiple instances from the same cohort of patients over a period of time, may assume that the errors for observations from the same subject correlate differently than errors between different patients. Fixed, random, and mixed effects models – core statistical methods within certain sub-disciplines in the sciences and social sciences – are forms of linear models adapted to handle applications such as resampled data.

Linear models also benefit from a strong theoretical background. The standard estimators, which we will explore in the following two sections, can be described in terms of weighted sums of the original data. Under weak assumptions, we can then draw on the central limit theorem and large sample theory to construct asymptotically valid confidence intervals and hypothesis testing frameworks. Importantly, most of this theory can be extended to the various extensions and complex assumptions often used in practice. Also, these theoretical tools are useful even when the primary task is one of prediction. Hypothesis tests aid in the process of deciding whether to add or delete a certain variable from a model. Confidence intervals, when combined with an estimate of the noise variance, are extensible to prediction intervals. These provide a range of likely values for newly observed data points, in addition to a singular 'best' value. We will see several ways in which these estimates are useful in practice when building predictive models.

The standard estimators for parameters in linear models can be calculated using relatively straightforward computational approaches. For this reason, linear models are often used in applications even when many of the aforementioned benefits do not directly apply. Notice that a linear model must be linear only relative to the β terms. If we have pairs of data (x_i, y_i) but believe that there is a non-linear relationship between x and y, we could build the model

$$y_i = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot x_i^2 + \cdots + \beta_p \cdot x_i^p + \epsilon_i. \tag{1.6}$$

Here it is difficult to discern a conceptual interpretation of each of the β_j terms. As a result, it is also hard to make use of confidence intervals and hypothesis tests

concerning them. However, the linear model framework is incredibly useful as it provides a computationally tractable way of estimating an arbitrarily complex relationship, by setting p as large as possible, between our two variables. Of course, the size of the dataset will limit the ultimate complexity of the model, but this is true regardless of the particular approach taken. We will expand at length on this variable expansion method in Chapters **??** and **??**.

References

LAB QUESTIONS

1. Here is a thing!