MTH 4230 Spring 2021
Module 2 Notes and Exercises

Linear Models

In an understated fashion typical of mathematical nomenclature, an unfathomably deep collection of models is referred to with the simple sounding “linear models.” Our textbook, in fact, is titled *Linear Models in R*, demonstrating the commonality of this broad term in statistical science.

In applied mathematics we also have, for example, the term “linear partial differential equations,” which refers to an unfathomably deep collection of deterministic equations for which exact solutions are rarely known. In MTH 4230, we are interested in families of data models with unknown parameter values and, per the objective of inferential statistics, our goal is to make a probabilistically correct statement about the unknown model parameters, usually in the form of an interval estimate or a hypothesis test conclusion. The values of the parameters determine the specific member of the family of models used for a particular data set.

In addition to mathematically precise inferential statistics procedures, we use more subjective statistical methods for statistical analysis, including descriptive statistics methods for organizing, graphing, and constructing informative numerical measures about multivariate samples. Further techniques for model diagnostics, validation, and selection require other types of imprecise analysis like pattern recognition and real world considerations.

In all but a few cases, the models we study are *linear with respect to the parameters.* Nonlinear transformations and combinations of variables are present in these models in other forms. Thus, “linear models” can be much more complicated than the simple name might imply; in particular, the response surface $\mu_Y = f(x_1, x_2, \ldots, x_n)$ need not be linear.

Specific Families of Linear Models. In the context of statistics, “linear models” is not a universally precise term. Precise names for specific families of linear models include the following:

- Simple (first-order) linear regression with one predictor variable
- Simple (first-order) linear regression with multiple variables
- Quadratic (second-order) linear regression with multiple variables
- Polynomial (nth-order) linear regression with multiple variables
- Linear regression with transformed variables
- General linear models, abbreviated GLM (with independent, normally distributed error terms)
- Generalized linear models, abbreviated GLiM or sometimes, confusingly, just GLM (error terms can be correlated or not normally distributed)

Within the ladder of generalizations indicated above, there are a number of special cases and mixed models that we look at more carefully; for example, models with qualitative predictors, autocorrelated time-series models, and models with qualitative responses. The latter two topics are GLiM’s but not GLM’s because the error is correlated and or non-normal.

Model Parameters. The modules our class is organized into use an approach that starts with the most restrictive, mathematically tractable models and culminates with the models from the most general families we consider. For example, our last module covers the logistic regression model that can be classified as a GLiM but not a GLM.

It is instructive to consider the most general families of models to start with to get an overview off where we are headed this semester.

The variables in the GLM form given below, denoted $X_1, \dots X_{p-1}$, do not need to represent different predictor variables. We use the symbol Y to represent the response variable. Each observation of actual data associated with the model is p-dimensional, denoted by

$$(X_{i1}, X_{i2}, \dots, X_{i,p-1}, Y_i)$$

In some studies each observation is called a *trial.*

The *General Linear Model (GLM)* is defined to be the $(p+1)$-parameter data model of the form,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i$$

where

- $\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$ is a dimension $n \times 1$ vector of response variable observations,
- $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}$ is a dimension $p \times 1$ vector of model parameter values,
- $\mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{pmatrix}$ is a dimension $n \times p$ matrix of predictor variable observations, and,
- $\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$ is a dimension $n \times 1$ vector of random numbers with a $N(0, \sigma^2)$ distribution. Note that σ is the $(p+1)^{\text{th}}$ parameter value for the GLM.

A consequence of the GLM model assumptions is that

$$\mu_i := E[Y_i] = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1}$$

More generally, the *Generalized Linear Model (GLiM)* is defined to be the $(p+1)$-parameter data model of a similar form as above, with the allowance that a so-called *link function g* relates the mean response variable, μ_i to the predictor variables according to the formula,

$$g(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1}$$

Further, the error terms ϵ_i do not need to be normally distributed or independent. This generalization also requires that the response variables Y_i have distributions that are from the *exponential family* of distributions, a very general family of probability distributions covered in statistical theory.

In our class, almost all models can be classified as GLMs; the time series and logistic regression models we cover at the end of the semester are GLiM's but not GLM's.

Data Collection.

Controlled Experiment Premise. In a controlled experiment, the experimenter controls the levels of the predictor variables and assigns a treatment, consisting of a combination of levels of predictor variables, to each experimental unit and observes the response [7]. The "control" comes from the fact that predictor variables are such that the experimenter can control their values.

Observational Study Premise. In a *confirmatory observational study*, non-experimental data that is observed are used to test (e.g confirm or not confirm) previous studies or beliefs. In an *exploratory observational study*, the researcher may have no pre-defined studies or beliefs that indicate which predictor variables are correlated with the response variable. For these studies, investigators consider several possible explanatory predictor variables that might be related to the response and search which ones might be most correlated with the response variable [7].

Model Identification. "All mathematical models are wrong, but some mathematical models are useful" [1].

As the quote implies, mathematical models are imperfect, so building a model can seem more like an art than a science, but general guidelines for the mathematical modeling process have been developed by several authors, including the sources cited below. The modeling strategy outlines below are different but have many similarities.

(1) In the reference [3], the author diagrams the following model-building process:
Step 1: Indicate the real-world situation being modeled. To get to Step 2, identify a limited number of quantities and their relationships to each other.
Step 2: Formulate a mathematical model. To get to Step 3, find a solution to the mathematical model.
Step 3: Form new relationships among the quantities that were not already identified in Step 1. To get to Step 4, evaluate the new relation for values of the variables not yet considered.
Step 4: Make predictions based on the new relations, and verify if the new relations make accurate predictions. Confirm or reject the new relations, return to Step 1, and repeat process as needed.

(2) In the reference [3], the author describes a model building process invented by Glenn Ledder:
Phase 1: Real World Situation (RWS). To go forward from RWS to CM, simplify and approximate with a conceptual model.
Phase 2: Conceptual Model (CM). Use mathematical derivations to go forward from CM to MM. Use model validation methods to go backwards from CM to RWS.
Phase 3: Mathematical Model (MM). Use mathematical analysis to go backwards from from MM to CM.

(3) In the reference [7], a strategy for building a regression model is given as follows:
Step 1: Collect data
Step 2: Preliminary checks on data quality
Step 3: Diagnostics for relationships and interactions
Decision: Are remedial measures needed? If so, implement measures and return to Step 3
Step 4: Determine several potentially useful subsets of predictor variables
Step 5: Investigate curvature and interaction effects more fully
Step 6: Study residuals and implement diagnostic tests
Decision: Are remedial measures needed? If so, implement measures and return to Step 5
Step 7: Select tentative model
Decision: Implement validity checks. If predictions produced by tentative model are not useful or valid, return to Step 1 after designing a new experiment
Step 8: Select and implement final regression model

In short, the strategy above can be reduced to the following four important phases:
Phase 1: Data collection and preparation
Phase 2: Reduction of predictor variables
Phase 3: Model refinement and selection
Phase 4: Model validation

(4) Finally, the Mathematical Contest in Modeling (MCM) is an international contest for high school students and college undergraduates. It challenges teams of students to clarify, analyze, and propose solutions to open-ended problems. The document [2], titled "20 Years of Good Advice," provides advice for teams competing in this modeling competition, including mathematical modeling strategies.

Exercise 1 (5 points): A prerequisite to this course is the concept of the best fit least squares regression lines for bivariate data sets. For this exercise, use a textbook, old homework assignment, or an OER elementary statistics book like *Introductory Statistics* by OpenStax.

Find a real-world bivariate linear regression problem from one of these sources that includes a scatterplot, calculation of the point estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, and calculation of the residual errors. Treating the point estimates as parameter values ($\beta_0 = \hat{\beta}_0$ and $\beta_1 = \hat{\beta}_1$), identify each of $\mathbf{Y}$, $\boldsymbol{\beta}$, $\mathbf{X}$, and $\boldsymbol{\epsilon}$ from the GLM form above.

Exercise 2 (5 points): Compare and contrast two of the four modeling strategies indicated in the Model Identification section of these notes. Which steps are similar? Which parts are different? How can you synthesize the two strategies?

Exercise 3 (10 points): Review past problems from the Mathematical Contest in Modeling that are posted online at the contest website:
https://www.comap.com/undergraduate/contests/mcm/previous-contests.php
Identify a past contest problem that interests you. Sketch a mathematical modeling approach that uses data analysis for some aspect of this problem. Include each of the following steps:

(1) What data collection strategies for the problem would you consider. Would you use a controlled experiment or an observational study approach?
(2) Identify several possible predictor variables of interest and a response variable of interest; how might you reduce this set of predictor variables for an initial analysis of a simple version of the model?
(3) What variable reductions should be considered to simplify the model?
(4) What data or predictions would you use to validate your model?

References

[1] Draper, N.R., Smith, H. *Applied regression analysis 3rd ed.* Wiley (1998)

[2] Driscoll, P., Griggs, J., Parker, M., Boisen, P., Fox, W.,Tortorella, M., Campbell, P. "20 Years of Good Advice" (.pdf). https://www.comap.com/undergraduate/contests/resources/PDF/20YearsofGoodAdvice.pdf

[3] Dunbar, S. *Mathematical Modeling in Economic and Finance: Probability, Stochastic Processes, and Differential Equations.* AMS/MAA Press (2019).

[4] Faraway, J. *Linear Models with R, 2nd ed.*. Chapman and Hall, (2014).

[5] Faraway, J. (April 27, 2018) *Linear Models with R* translated to Python. Blog post. https://julianfaraway.github.io/post/linear-models-with-r-translated-to-python/

[6] Fahrmeir, Kneib, Lang, Marx , *Regression.* Springer-Verlag Berlin Heidelberg (2013).

[7] Kutner, Nachtsheim, Neter, Li *Applied Linear Statistical Models, 5th ed.* McGraw-Hill/Irwin (2005).