MTH 4230 Spring 2021
Module 8 Notes and Exercises (Project 7)

Forms of the General Linear Model

In the Module 7 notes, the matrix form of the multiple linear regression data model for a set of response variable values $\mathbf{Y}$ as a function of a set of vector-valued predictor variable $\mathbf{x} = (x_1, \ldots, x_{p-1})$ is given by

$$\mathbf{Y} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{X}$, $\boldsymbol{\beta}$, $\boldsymbol{\epsilon}$ and $\mathbf{Y}$ are as indicated in the Module 6 notes and $E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$.

Although the form of this model is *linear in the parameters*, countless nonlinear transformations of the variables of direct interest to the researcher can be used to generalize the model in ways that preserve the linear form of the model. The response variable Y can be related to a response variable of direct interest by way of transformations. Predictor variables can also be (possibly multivariate) transformations of other predictor variables of direct interest. There are no explicit restrictions on conditional distributions between two or more of the predictor variables.

Here, we review some specific subfamilies of the general linear model that do fit within the framework of the model; later, in Modules 12, 13 and 14, we review some *generalized* linear models that are important in applications but *do not* fit with the framework of the general linear model.

Notation note: To simplify the indices below, we omit the i subscript indicating individual experimental units.

Regression on transformed predictors. Often a transformation applied to the response variable or a single predictor variable of direct interest can be used to put a real world data model into the GLM form.

Two examples of underlying data models that can be put into the GLM form using transformation of the response variable follow (what transformation?):

$$Y = Ce^{\beta_1 X_1 + \beta_2 X_2 + \epsilon} \qquad\qquad Y = \frac{1}{\beta_1 X_1 + \beta_2 X_2 + \epsilon}$$

Transformations of predictors are commonly applied in practice too. Two examples of underlying data models that can be put into the GLM form using transformation of the response variable follow (what transformation?):

$$Y = \beta_0 + \beta_1 \ln(X_1) + \frac{\beta_2}{\sqrt{1 + X_2}} + \epsilon \qquad\qquad Y = a(x - x_0)^2 - d$$

Polynomial regression. Polynomials involving one or more variables of interest can be incorporated into the GLM by encoding higher order powers in the polynomial with additional model variables. For example, a quadratic regression model on the single variable X can be written,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

A more complicated third-order, multivariable model with predictor variables X, U, and V is, for example,

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon \\ &= \beta_0 + \beta_1 X + \beta_2 U + \beta_3 V + \beta_4 XV + \beta_4 XV^2 + \epsilon \end{aligned}$$

Once it is clear that these polynomial models have the form of a general linear model, we rarely bother with the notational substitutions indicated above, and simply express the model in terms of polynomial combinations of predictor variables denoted by $X_1, X_2, \ldots X_n$. We can also use a standardized notation for the β coefficients, and introduce the substitution $x_j = X_j - \overline{X_j}$ (to avoid strong correlations between predictors). Note that the fitted values and residuals in terms of X will be the same as those of x.

Using this notation, the **single-predictor, nth order polynomial regression model** has the general form,

$$Y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \cdots + \beta_{11\cdots 1} x^n + \epsilon$$

where the final β subscript has n copies of 1 and $x = X - \overline{X}$

The **two-predictor, second order polynomial regression model** has the general form,

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon$$

where $x_j = X_j - \overline{X_j}$.

The **three-predictor, second order polynomial regression model** has the general form,

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \epsilon$$

where $x_j = X_j - \overline{X_j}$.

Note that a researcher's hypothesized real-world form of a model may not appear to match the form of a polynomial (or other) regression model, but algebraic operations can often be used to put the model in a standard form. For example, how can the model below be put in a polynomial regression form?

$$Y = a(x - x_0)^2 - d + \epsilon$$

Regression on qualitative predictors. Qualitative and binary response variables are not accommodated by the General Linear Model family (rather, *generalized* linear models are used in this case), but qualitative and binary predictor variables can be used in the general linear model.

For a basic example consider a GLM with three predictor variables, X_1 a real-valued, continuous predictor variable, X_2 a binary predictor variable with levels $\{0, 1\}$, and U, a categorical predictor variable with levels $\{A, B, C\}$. We can encode a general linear model for response variable Y with respect to these three predictors as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon$$

where,

$$X_3 = \begin{cases} 1 & U \text{ is at level } A \\ 0 & \text{otherwise} \end{cases} \qquad X_4 = \begin{cases} 1 & U \text{ is at level } B \\ 0 & \text{otherwise} \end{cases} \qquad X_5 = \begin{cases} 1 & U \text{ is at level } C \\ 0 & \text{otherwise} \end{cases}$$

When binary variables are involved, interaction effects have a specific interpretation. For example consider the model above with only the continuous variable, X_1, and the binary variable, X_2. Given that the $X_2 = 0$, the model becomes,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 \cdot 0 + \epsilon = \beta_0 + \beta_1 X_1 + \epsilon$$

conversely, if given that $X_2 = 1$, we have that

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 \cdot 1 + \epsilon = (\beta_0 + \beta_2) + \beta_1 X_1 + \epsilon$$

Thus, for this additive model, the effect of different levels of the binary variable X_2 is a constant shift of the linear relationship between Y and X_1.

On the other hand, if we include an interaction term, the value of the binary predictor has a more dynamic effect on the model. The interaction term gives the model the form,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

We again find that, given $X_2 = 0$,

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

but now we find that, given $X_2 = 1$,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 + \beta_3 X_1 + \epsilon = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 + \epsilon$$

In this case, both the intercept and the slope of the linear relationship between Y and X_1 change with respect to the level of X_2.

Project 7

Project 7 is a group exercise, but each individual student will submit their own project report for assessment. The project presentations, presented in groups, is scheduled for 4/19/21 and 4/21/21.

Exercise (20 points): Complete Textbook Exercise 1, 2 **or** 3 from Chapter 12 (page 195). Work with a team of at least two other students in the course (3-5 students per team). Adhere to the textbook's instructions: "A full answer requires you to perform a complete analysis of the data including an initial data analysis, regression diagnostics, a search for possible transformations and a consideration of model selection. A report on your analysis needs to be selective in its content. You should include enough information for the steps leading to your selection of model to be clear and reproducible by the reader. But you should not include everything you tried. Dead ends can be reported in passing but do not need to be described in full detail unless they contain some message of interest. Above all your analysis should have a clear statement of the conclusion of your analysis."

Each group is expected to present their analysis in class using professional presentation software (e.g. Powerpoint, LaTeX Beamer, R Markdown). Every individual student is required to submit a project report with each of the following items:

(1) A precise description of the role and mathematical contribution of each member of the group. An indication of which components of the final presentation are completed by which group members. *This item should be identical for all members of the same group, the proceeding items, below, are individualized.*
(2) An excerpt of slides that represent your individual contribution to the project presentation done in class.
(3) An indication of the statistical software commands and outputs used to create your individual contribution to the group project presentation.
(4) Your description of the final conclusions of the analysis in terms of the real world problem.

References

[1] Cornillon, Pierre-Andre. *R for Statistics, 1rst ed.*. Chapman and Hall, (2012).
[2] Draper, N.R., Smith, H. *Applied regression analysis 3rd ed.* Wiley (1998)
[3] Faraway, J. *Linear Models with R, 2nd ed.*. Chapman and Hall, (2014).
[4] 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/
[5] Fahrmeir, Kneib, Lang, Marx , *Regression.* Springer-Verlag Berlin Heidelberg (2013).
[6] Kutner, Nachtsheim, Neter, Li *Applied Linear Statistical Models, 5th ed.* McGraw-Hill/Irwin (2005).