MTH 4230 Spring 2021
Module 12-14 Notes and Exercises (Project 9)

Generalized Linear Models

While the General Linear Model (GLM) introduced in previous models can accommodate complex models with nonlinear response surfaces and complicated interactions between the variables of interest, there are still a number of restrictive assumptions that prevent us from using a GLM to model certain real world systems in a correct way.

In this module, we survey models that are in the GLM family but can be classified as special cases of a more general family of models called *Generalized Linear Models*. All the models covered in these notes can be understood as special cases of Generalized Linear Models.

The *Generalized Linear Model* family for a single response variable Y is defined as follows.

- Assume the response variable observations Y_i are independent random variables with finite means μ_i with a distribution from the *exponential family* of random variables covered in statistical theory.
- A linear combination of predictor variables, similar to the right side of the GLM, is defined. Denote these multivariate linear combinations of observations by
$$\mathbf{X}_i\beta = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_{p-1} x_{i(p-1)}$$
- A *link function*, g, that satisfies certain technical constraints relates the linear combination of the predictors to the response variable means, $\mathbf{X}_i\beta = g(\mu_i)$. Non-constant variances of the response variables are allowed, but they must all be a function of the μ_i associated with the predictors through the link function.

This formal conceptualization of the Generalized Linear Model family is very technical and not particularly helpful for practitioners of statistical applications. On a practical level, there are four possible violations of the GLM assumptions that Generalized Linear Models are used to accommodate:

(1) The model is fundamentally nonlinear with respect to the parameters, not just a linear function that relates transformations of the original variables of interest.
(2) The error terms do not have a constant variance.
(3) The error terms are not independent.
(4) The error terms do not have a normal distribution.

We use the enumerated list above when indicating the utility of specific families of Generalized Linear Models that are not also GLMs. The remainder of these notes is a survey of specific subfamilies of the Generalized Linear Model that are common in applications.

Logistic Regression. Notation comment: In many books, π or π_i is used to symbolize $E[Y]$ or $E[Y_i]$, respectively, in the logistic models below; we avoid this notation and just use $E[Y]$ or $E[Y_i]$.

Technically, a response variable without a continuous range of possible values cannot possibly satisfy the GLM assumptions because the response, Y, in the GLM is a continuous, normally-distributed random variable. In cases where the response is discrete, but has a large number of possible values (e.g. the sizes of ant colonies), a continuous random variable

might be a suitable approximation of the response. If the response is binary, or has relatively few possible values, it cannot be reasonably approximated by a normal distribution and a different model is required.

In the case of a binary response variable, a common choice for modeling Bernoulli (random, $\{0, 1\}$-valued) response variables is called the logistic regression model. The output of the logistic regression family of models is interpreted, indirectly, as a likelihood that the binary response variable has value 1. All of the practical assumptions for the GLM except for (3), the independence of error terms, are violated in the logistic regression model.

The simple logistic regression model for the likelihood, $E[Y]$, that the Bernoulli random variable, Y, is equal to 1 is given by

$$Y = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)} + \epsilon = E[Y] + \epsilon$$

Here, for each $i = 1, \ldots, n$, the distribution of ϵ_i is given by $P\left[\epsilon_i = 1 - E[Y_i]\right] = E[Y_i] = 1 - P\left[\epsilon_i = -E[Y_i]\right]$. If the observations are random, the response is viewed as a conditional mean given the levels of the predictors X_i. If we express this in terms of an observed set of experimental units of size n, we have

$$Y_i = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} + \epsilon_i = E[Y_i] + \epsilon_i \qquad \text{for } \; i = 1, \ldots, n$$

The ϵ_i are assumed to be independent, but are not normally or identically distributed.

We omit the i subscript in most of the expressions in these notes for the sake of decluttering the notation, but we will continue to avoid using i and n to represent any other objects in the model. Note that X can be a transformation of another variable of direct interest to the researcher and the response Y can be easily transformed to any other binary random variable with a linear map.

Generalizing the model above, the multiple logistic regression model for the likelihood, $E[Y]$, that the Bernoulli random variable, Y, is equal to 1 is given by

$$\begin{aligned} Y &= \frac{\exp(\mathbf{X}\beta)}{1 + \exp(\mathbf{X}\beta)} + \epsilon \\ &:= \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1})}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1})} + \epsilon \\ &:= E[Y] + \epsilon \end{aligned}$$

where the distribution of ϵ is defined in the same way as in the single predictor case. Note that $\mathbf{X}$ is a vector of predictor values in this context and not a model matrix like in previous module notes.

Given a real world set of predictor observations and response observations, estimates of the parameters $\beta_0, \ldots, \beta_{p-1}$ are computed using numerical procedures that maximize a theoretical likelihood function for the model. The estimated parameter values are denoted $\mathbf{b} = (b_0, \ldots b_{p-1})$.

To recover a practical interpretation of the model from the estimated coefficients, we can compute

$$\widehat{E[Y_i]} = \frac{\exp(\mathbf{Xb})}{1+\exp(\mathbf{Xb})}$$

which is interpreted is the estimated probability that $Y = 1$ given a that the predictor values are the elements of the vector $\mathbf{X}$.

Derivations of the maximum likelihood estimates of parameters in the models can be onerous, and practitioners rely on computational methods to produces estimates and test hypotheses associated with these models. Comparing models is similar is some respects to the GLM case: Model selection criteria like AIC are still well-defined and can be minimized to produce more effective models.

The interpretation of the coefficients in the model are a little bit more complicated than before. For simple logistic regression, the follow can be shown to be a correct interpretation of the estimate of b_1 produced computationally: b_1 is a maximum likelihood estimated of β_1 and $\exp(b_1)$ is the estimated change in the odds ratio of the event $Y = 1$ compared to the event $Y = 0$ when the predictor variable changes by one unit. The interpretation is the same for multiple logistic regression, except that $\exp(b_j)$ for $j = 1, \ldots, p-1$ is the change in the odds ratio with respect to unit change in predictor j if all other predictors are held constant.

The theoretical motivation and derivation for confidence intervals and hypothesis tests associated with multiple logistic regression models are very involved, but computational methods in R can be used to construct these intervals and execute hypothesis test similar to those associated with the GLM. Further, the AIC and BIC procedures for ranking models are still defined and used in practice for variable selection in multiple logistic regression applications.

Diagnostic plots can be difficult to interpret in logistic regression applications, and outliers should not exist since the response only has two possible values. Residuals are typically transformed (e.g. Pearson residuals) and plots can be used to check model accuracy but not much else since the errors are not assumed to be identically distributed or have constant variance.

Poisson Regression. As already indicated at the start of the previous section, if a response variable, Y, is binary, or has relatively few possible values, it cannot be reasonably approximated by a normal distribution and a different model is required.

In the case of a response variable with values in the natural numbers, a common choice for modeling the response variables is called the Poisson regression model. The Poisson random variable is one of the most practically important random variables, and the Poisson regression model uses a Poisson random variable to model the response variable Y.

As with the logistic regression model all the practical assumptions for the GLM except for (iii), the independence of error terms, are violated in the Poisson regression model. The Poisson regression model can be stated succinctly as follows.

Let Y_i be independent Poisson random variables with expected value μ_i so that,

$$P[Y_i = j] = \frac{\mu_i^j \exp(-\mu_i)}{j!} \qquad \text{for} \quad j = 0, 1, 2, 3, \ldots$$

Further assume that the values μ_i are determined by a function of the predictor variable vectors $\mathbf{X}_i$ and the model coefficients, $\beta_0, \beta_1, \ldots, \beta_{p-1}$, say $\mu_i = g(\mathbf{X}_i, \beta_0, \beta_1, \ldots, \beta_{p-1})$. Then the generalized linear model of the form

$$Y_i = \mu_i + \epsilon_i$$

where ϵ_i, the error term, is determined by the fact that Y_i is Poisson with mean μ_i. Note that, by construction, $E[\epsilon_i] = 0$.

Now, in general the form of $\mu_i = g(\mathbf{X}_i, \beta)$ can take on many forms, but the outputs of g must be positive. Typical choices of g include,

$$\begin{aligned} \mu_i = g(\mathbf{X}_i, \beta) &= \mathbf{X}_i\beta \\ \mu_i = g(\mathbf{X}_i, \beta) &= \exp(\mathbf{X}_i\beta) \\ \mu_i = g(\mathbf{X}_i, \beta) &= \ln(\mathbf{X}_i\beta) \end{aligned}$$

with the second item in the list above being the most commonly used response in polynomial regression models.

As with logistic models, inferential methods using maximum likelihood estimates and their distributions rely on computational techniques. Model selection criterion like the AIC can be used to select effective models and the interpretation of model coefficients in real world terms must be done with care. Researchers are often most interested in estimating mean responses or probabilities of individual responses associated with specified levels of the predictor variables.

Time Series Models. *Time Series Models* are a family of statistical models where observations are indexed by time or other real-world sequential component (e.g. number of rounds of a card game). Notably, a time variable, t, need not be explicitly included in these models. In the *autonomous* examples below, time only enters the model as an index, but the correlations of the response and predictors with respect to time introduce autocorrelations in the error terms that are not accommodated by GLMs.

The main practical assumption associated with the GLM that is violated is (3) since the errors are non-independent. Many important time series models have non-constant variance, but the models we consider below do in fact have a constant error variance and normally distributed error.

The foundations of time series models are properly covered by the theory of stochastic processes; discrete stochastic processes use a discrete time index and continuous stochastic processes use a continuous time index. Time series are just one small class of real world stochastic models, which can also be indexed by spatial or spatio-temporal structures. The formalization of certain time series models as members of the *generalized linear model* family is useful and allows researchers of time series models alternative perspectives.

AR(1) *Model.* In this subsection we exemplify one specific class of models called first-order autoregressive models, or $AR(1)$ models. These discrete-time (minutes or days or years, ect...) models are important building blocks in climatology, meteorology, hydrology, biology, and a range of other applications. The $AR(1)$ is defined by,

$$Y_t = \beta_0 + \beta_1 X_{t,1} + \beta_2 X_{t,2} + \cdots + \beta_{p-1} X_{t,p-1} + \epsilon_t$$

where ϵ_t are non-independent, 0-mean error terms related by the equation

$$\epsilon_t = \rho\epsilon_{t-1} + z_t$$

and z_t are independent, normally-distributed random variables with mean 0 and variance σ^2.

This model has $p+2$ parameters because the parameter ρ (greek letter rho), which satisfies $-1 < \rho < 1$, is an additional parameter to the GLM parameters called the *autocorrelation coefficient*.

$AR(1)$ Basic Results. One finds that the variance of the error terms in this models is not σ^2 but is still constant, in fact,

$$\sigma^2_{\epsilon_t} = \frac{\sigma^2}{1-\rho^2}$$

for all t. Further the covariance between subsequent error terms can be derived as

$$COV(\epsilon_t, \epsilon_{t-1}) = \frac{\rho\sigma^2}{1-\rho^2}$$

and the covariance between subsequent error terms that are s terms apart can be derived as

$$COV(\epsilon_t, \epsilon_{t-s}) = \frac{\rho^s\sigma^2}{1-\rho^2}$$

This is called the auto-covariance of the process and it decays to 0 as $s \to \infty$ since $|\rho| < 1$. Notably, the population variance-covariance matrix associated with the error terms (see Module 6 notes, page 3) cannot be expressed in the diagonal form of the GLM model given by $\boldsymbol{\sigma^2} = \sigma^2\mathbf{I}$; in fact, the diagonal terms of the matrix $\boldsymbol{\sigma^2}$ for the $AR(1)$ model are all equivalently $\frac{\sigma^2}{1-\rho^2}$ and the off-diagonal terms are nonzero, given by powers of ρ.

Tests for Autocorrelation. The Durbin-Watson test for autocorrelation can be viewed from the GLM perspective as a test for if, viewed as an $AR(1)$ model, the model is a GLM or an $AR(1)$ model that is not a GLM. The test proceeds as follows.

Durbin-Watson test for $\rho = 0$

- **Null Hypothesis** $H_0 : \rho = 0$
- **Alternative Hypothesis** $H_A : \rho \neq 0$
- **Test Statistic**

$$D = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

- **P-value**

  The P-value associated with the Durbin-Watson test statistic is very difficult to compute. Historically, a range of test statistics led to inconclusive results. Computation of the P-value relies on bootstrapping methods that can be unreliable for small data sets.

A computational exposition of the Durbin-Watson test using R is presented in Subsection 6.1.3 of the textbook *Linear Models in R*

Applications to Lag Effects in Time-series. As in Section 4.3, of the textbook *Linear Models in R*, the *lagged variable* $AR(1)$ *process* has the following general form

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \epsilon_t$$

where ϵ_t are non-independent, 0-mean error terms related by the equation

$$\epsilon_t = \rho \epsilon_{t-1} + z_t$$

and z_t are independent, normally-distributed random variables with mean 0 and variance σ^2.

The textbook offers a specific example of an $AR(3)$ process with three lagged variables. In the textbook example, the time-series data is log-transformed and converted into the form of a data frame with lag terms in separate columns in order to apply the model.

Nonparametric Models. In many important real world applications, all of the practical assumptions associated with the GLM listed above are satisfied except for (4), normality of the error term distribution. In some additional cases (2) and (4) are violated, meaning non-constant variance of error is also possible.

Remedial methods that do not dramatically alter the GLM structure can often be employed in these cases. This allows for employing similar real world analysis of the models at the cost of some mathematical adjustments that allow us to employ the same confidence estimates and inferential methods. We surveys three specific techniques.

Bootstrap Sampling. Bootstrapping is a term that has different precise meanings in different contexts (e.g. statistics and mathematical analysts use the term for different things). In the context of linear models, bootstrap sampling is a the use of resampling (with replacement) of the original sample to infer the true variance of a maximum likelihood estimator being used to build interval estimates of model parameters.

Bootstrap methods are broad and research is ongoing, but several specific bootstrap sample approaches have been adopted for linear models, especially models with error terms that have non-normal distributions. Section 3.3 (Permutation Tests) and 3.6 of the textbook, *Linear Models in R*, covers bootstrap confidence interval used in linear regression and demonstrates their utility for non-normal error distributions.

Weighted Least Squares. Weighted least squares uses a set of weights, $1/w_j$ for $j = 1, \ldots, n$ in the calculation of the linear model, essentially building a linear model on the pairs $(\sqrt{w_i}x_i, \sqrt{w_i}y_i)$ instead of the original data. Residuals are also weighted as $\sqrt{w_i}\epsilon_i$ in the model equation.

Including weights allows one to transform data with pronounced non-constant variance in the error distribution into a model that satisfies the GLM constant variance assumption. In many important cases, the appropriate weights, w_j, to use in the transformation are unclear and also have to be estimated. In these cases, the mathematical model becomes even more complicated and most practical applications rely on computational methods.

Section 8.2 of the textbook, *Linear Models in R*, covers weighted least squares methods used in linear regression and demonstrates the use of the `gls` command in *R* for implementation of the weighted least squares methods.

Project 9 Exercises

Exercise 1 (5 points): Consider the `spector` data set in the `faraway` package. Use a logistic regression model to model `grade` as a function of all other variables.

(1) Identify the sub-model that still includes `psi` as a predictor and minimizes the AIC for this logistic regression model.
(2) Identify the sub-model that still includes `psi` as a predictor and minimizes the standard error associated with the estimate of the coefficient for `psi`.
(3) Interpret the `psi` coefficient from the sub-models from the first two parts in terms of real-world odds ratios.

Exercise 2 (5 points): Consider the `ships` data set in the `MASS` package. Use a Poisson regression model to model `incident` as a function of all other variables.

(1) Identify the sub-model that minimizes the AIC for this logistic regression model.
(2) Now consider all possible interactions of predictors in your Poisson regression model. Does the inclusion of the interaction term lower the AIC below the model chosen in the previous part?
(3) Identify the sub-model of the interactions model consider in the previous part that minimizes the AIC for this logistic regression model.

Exercise 3 (5 points): Consider the `globwarm` data set. Use the response variable `nhtemp` and select three proxy variables as predictors.

(1) Using the data from 1856 and on, build a GLM using the three proxy variables as predictors. Use a Durban Watson test to test if $\rho = 0$.
(2) Using the data from 1856 and on, build a lag 1 autoregressive model using only previous year's nhtemp as a predictor. Complete this using code similar to Faraway's code on page 55. Use a Durban-Watson test to test if $\rho = 0$.

Exercise 4 (5 points): Complete Textbook Exercise 4 from Chapter 4 (page 57), stated as follows: The dataset `mdeaths` reports the number of deaths from lung diseases for men in the UK from 1974 to 1979.

(1) Make an appropriate plot of the data. At what time of year are deaths most likely to occur?
(2) Fit an autoregressive model of the same form used for the airline data in Section 4.3. Are all the predictors statistically significant?
(3) Use the model to predict the number of deaths in January 1980 along with a 95% prediction interval.

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).
[7] Nelder and Wedderburn. "Generalized Linear Models," *Journal of the Royal Statistical Society A*, 135 (1972), pp.370-384.