Lecture 8: More on GLMs

(Text Sections 3.1, 3.4, 3.5)

We have seen in Lecture 7 that there are three components to a GLM (the distribution of Y_i, the linear predictor, and the link function).

Formally, there are three necessary and sufficient conditions for a model to be a GLM:

1. The Y_i's are independent.
2. The distribution of Y_i is in the exponential family in canonical form
3. The link function $g(\mu_i)$ is linear in the regression coefficients ($\boldsymbol{\beta}$), where g is monotonic and differentiable over the range of μ_i.

Example: Multiple sclerosis data

In a study to investigate the effect of the drug interferon on MS, patients were randomly assigned to one of three treatment groups (Placebo, Low dose, High dose). Each month for a period of 2 years the patients received MRI scans, and the number of lesions observed on each was recorded. A question arose surrounding the randomization process, and it was decided to test whether the mean number of lesions was equal in all three treatment groups at baseline. Let Y_i be the observed lesion count on the i^{th} patient at baseline. Furthermore, let

$$x_{i2} = \begin{cases} 1, & i^{th} \text{ patient is in LD group} \\ 0, & 0 \text{ otherwise} \end{cases},$$

and let

$$x_{i3} = \begin{cases} 1, & i^{th} \text{ patient is in HD group} \\ 0, & 0 \text{ otherwise} \end{cases}.$$

Since our response is a count, a natural preliminary choice for the distribution of Y_i is Poisson(μ_i). We specify the linear predictor as $\eta_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}$.

To specify the link function, recall that the Poisson distribution with mean μ has the canonical form with natural parameter $\log \mu$. Since the range of the log function is the real line, the *log link* function is natural for Poisson data:

$$g(\mu_i) = \log \mu_i = \eta_i = \sum_{j=1}^{3} x_{ij}\beta_j.$$

To derive the likelihood, we first invert the link function to compute

$$\mu_i = g^{-1}(\eta_i) = \exp\left\{\sum_{j=1}^{3} x_{ij}\beta_j\right\}.$$

The log-likelihood is then

$$\begin{aligned}\log \mathcal{L}(\boldsymbol{\beta};\mathbf{y}) &= -\sum_{i=1}^{n}\mu_i + \sum_{i=1}^{n} y_i \log \mu_i - \sum_{i=1}^{n}\log y_i! \\ &= -\sum_{i=1}^{n} e^{\sum_{j=1}^{3} x_{ij}\beta_j} + \sum_{i=1}^{n} y_i \sum_{j=1}^{3} x_{ij}\beta_j - \sum_{i=1}^{n}\log y_i!\end{aligned}$$

Again, the MLE of $\boldsymbol{\beta}$ does not have a closed form.

Q: Would this model be appropriate to model all the lesion counts (i.e. collected over time)?

Q: How do we interpret β_2?

Q: Why couldn't we simply transform the lesion counts (e.g. by taking the log of each count) to reduce the skewness in the data and then use the normal approximation? Note: There are many 0's in this data set.

It is important to distinguish between the roles of the μ_i's and $\boldsymbol{\beta}$ in the model. We will address this issue in two cases: the case where we have replicates (i.e. multiple responses observed at the same covariate values), and the case where we have no replicates (i.e. each response is associated with a different combination of covariate values).

Case I: No replicates

Consider the simple problem of modelling a normally distributed variable, Y_i, with constant variance σ^2 in the presence of a continuous predictor variable, x_i. In this case, we are assuming that the values of x_i are unique (i.e. that there are no replicates). The most general model for Y_i actually doesn't contain x_i at all. This model is called the *saturated model*, and is given by

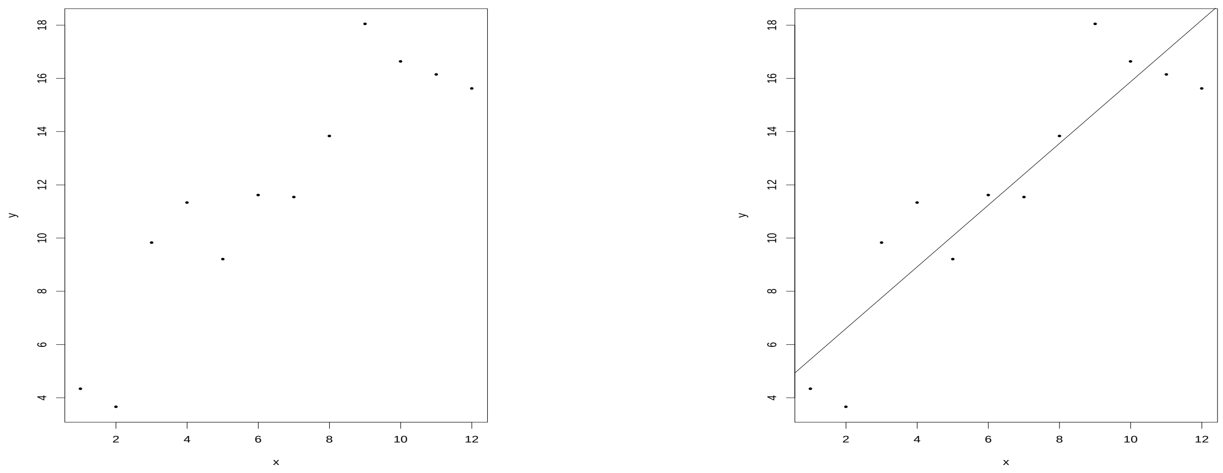
$$Y_i \sim N(\mu_i, \sigma^2),$$

$i = 1, \ldots, n$, where the μ_i's are the unknown parameters to be estimated. In contrast, the *simple linear regression model* would be

$$Y_i \sim N(\beta_1 + \beta_2 x_i, \sigma^2),$$

$i = 1, \ldots, n$, where β_1 and β_2 are the unknown parameters. Figures 1a and 1b show plots of the data along with the predicted values of Y_i for the two models. Note that the saturated model fits the data exactly.

Figure 1: Plot of unreplicated data and the fitted models.



a) Estimated saturated model

b) Estimated regression model

These ideas carry over to the GLM setting as well. In the saturated model, we allow the μ_i's to be unrestricted. In other words, we would allow a different mean value for each Y_i, unconstrained by $\boldsymbol{\beta}$. We treat the μ_i's as the parameters to be estimated. Although this model is very flexible, it is not particularly useful. Specifically, such models often lead to poor predictions of Y (for a given value of x). In addition, we cannot estimate the standard deviation of $\hat{\mu}_i$, which limits our ability to make inferences.

Typically, we would instead use a model where μ_i is parameterized in terms of $\boldsymbol{\beta}$ (the regression approach). This is the role of the link function and linear predictor:

$$g(\mu_i) = \sum_{j=1}^{p} x_{ij}\beta_j.$$

This model is less general than the saturated model, but is parsimonious, and clearly defines the relationship between Y_i and x_{ij}. In this model, the β_j's are treated as the unknown parameters to be estimated, not the μ_i's. If we wish to estimate μ_i, we simply invert the link function and plug in the MLEs of $\boldsymbol{\beta}$:

$$\hat{\mu}_i = g^{-1}\left(\sum_{j=1}^{p} x_{ij}\hat{\beta}_j\right).$$

We will see later how to compute a standard error for this quantity.

Case II: Replicates

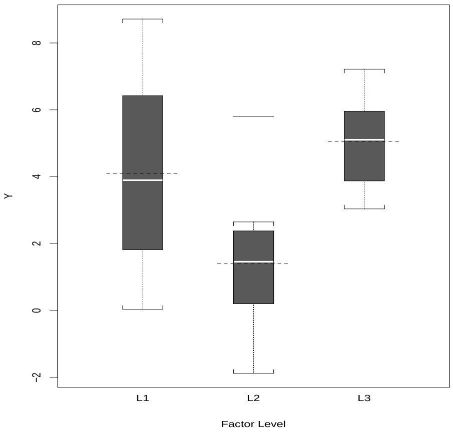
Similar ideas apply in the case where there are replicates. In this case, there are multiple responses collected for given values of the covariates. Replicates often occur when the predictor variables are categorical (e.g. in the ANOVA context). However, we can also obtain replicates when the predictor variables are continuous. (You may recall the concept of "pure error" estimation from previous classes.)

For simplicity, let's consider the case where there is one categorical predictor variable with m levels, and the response is normally distributed. Define Y_{hi} as the i^{th} response observed at the h^{th} level of the predictor. Then, the saturated model is

$$Y_{hi} \sim N(\mu_h, \sigma^2),$$

$i = 1, \ldots, n$, $h = 1, \ldots, m$, and the μ_h's are the parameters to be estimated. Figure 2 shows a plot of the data along with the predicted values of μ_h for each h ($m = 3$).

Figure 2: Boxplot of replicated data. The broken lines indicate the $\hat{\mu}_h$'s.



In this case, the saturated model is the same as the familiar ANOVA model. Note that, in general, it would not be reasonable to use a regression model to describe the relationship between μ_h and the predictor. (A regression model assumes a functional relationship between μ_h and the predictor variable, such as a straight line.)

The saturated model in the GLM setting would be expressed as

$$g(\mu_h) = \theta_h,$$

where Y_{hi} has some distribution in the exponential family with mean μ_h. Here, the θ_h's are the parameters to be estimated. Note that there is a 1-1 correspondence between θ_h and μ_h.

In the case where the predictors variables are continuous and there are replicates, we have the choice of fitting a regression-type model (as described in Case I) or an ANOVA-type model (as described for categorical predictors). The former is typically more parsimonious, while the latter has the advantage of not assuming a specific functional form for the relationship between μ_{hi} and the predictors.