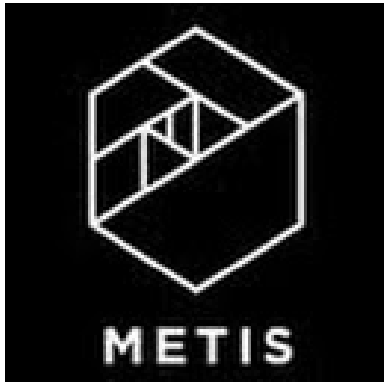
Generalized Linear Models

Luc Demortier

Metis, New York, August 1 2017



Ordinary Linear Models

Recall Ordinary Linear Regression (Ordinary Least Squares):

$$Y_i \;=\; \beta_0 + \beta_1\, X_{1i} + \beta_2\, X_{2i} + \ldots + \beta_p\, X_{pi} + \epsilon_i \qquad \text{for} \qquad i = 1, \ldots, n.$$

The classical assumptions for this model are:

1. Linearity in the parameters;
2. Identifiability: no collinearities;
3. Fixed covariates (X_i's), or $\mathbb{C}\mathrm{ov}(X_i, \epsilon_i) = 0$;
4. More observations than parameters;
5. Sufficient range in the X values;
6. Normally distributed errors ϵ_i;
7. Mean error is zero: $\mathbb{E}(\epsilon_i) = 0$ for all i;
8. Homoskedasticity: $\mathbb{V}\mathrm{ar}(\epsilon_i) = \sigma^2$ for all i;
9. No serial correlation: $\mathbb{C}\mathrm{ov}(\epsilon_i, \epsilon_j) = 0$ for any $i \neq j$;
10. The model is correctly specified.

Ordinary Linear Models

Recall Ordinary Linear Regression (Ordinary Least Squares):

$$Y_i \;=\; \beta_0 + \beta_1\, X_{1i} + \beta_2\, X_{2i} + \ldots + \beta_p\, X_{pi} + \epsilon_i \qquad \text{for} \qquad i = 1, \ldots, n.$$

The classical assumptions for this model are:

1. Linearity in the parameters;
2. Identifiability: no collinearities;
3. Fixed covariates (X_i's), or $\mathbb{C}\mathrm{ov}(X_i, \epsilon_i) = 0$;
4. More observations than parameters;
5. Sufficient range in the X values;
6. Normally distributed errors ϵ_i;
7. Mean error is zero: $\mathbb{E}(\epsilon_i) = 0$ for all i;
8. Homoskedasticity: $\mathbb{V}\mathrm{ar}(\epsilon_i) = \sigma^2$ for all i;
9. No serial correlation: $\mathbb{C}\mathrm{ov}(\epsilon_i, \epsilon_j) = 0$ for any $i \neq j$;
10. The model is correctly specified.

Generalized Linear Models

Generalized linear models allow one to consider nonnormal response variables with a nonlinear relationship to the predictors. They have three components:

1. The Random Component: This is the response variable Y_i and its distribution, which does not need to be normal. It can be binomial, multinomial, Poisson, negative binomial, ...
2. The Systematic Component: A linear combination of features, also known as linear predictor:

$$\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi}$$

3. The Link Function: A smooth and invertible function g, usually nonlinear, that connects the random and systematic components, and removes restrictions on the range of the systematic component:

$$g(\mu) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi},$$

where:

$$\mu = \mathbb{E}(Y_i \mid X_i).$$

Examples of Generalized Linear Models

Random Component	Link Function	Systematic Component	Model
Normal	Identity	Continuous	Linear
Binomial	Logit	Mixed	Logistic
Poisson	Log	Mixed	Loglinear

The link functions are defined as follows:

$$\begin{aligned} \text{Identity:} \quad & g(\mu) = \mu \\ \text{Logit:} \quad & g(\mu) = \log\left(\frac{\mu}{1-\mu}\right) \\ \text{Log:} \quad & g(\mu) = \log(\mu) \end{aligned}$$

There are other link functions, and other possible combinations of random components and link functions!

GLMs versus Response-Variable Transformations

A technique that is sometimes used in linear regression is to transform the response variable before fitting. How is this different from GLM?

- A response-variable transformation must meet two objectives:
  1. It must linearize the regression of the response on the predictors;
  2. It must "normalize" the distribution of the response.

  In contrast, GLMs separate the choice of transformation from that of the response distribution (within some constraints).

- Another difference is that with a transformation of the response we are modeling the expectation of the transformed response:

$$\mathbb{E}\left[g(Y_i) \mid X_i\right] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi},$$

  whereas with GLMs we are modeling the transformed expectation of the response:

$$g\left[\mathbb{E}(Y_i \mid X_i)\right] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi}.$$

  This is not the same thing when the link function is nonlinear.

The Logistic Model

Let's start with an exterminator example. Beetles were exposed to gaseous carbon disulphide at various concentrations (in mg/L) for 5 hours and the numbers of beetles killed were noted:

Dose	Exposed	Killed	Proportion
49.1	59	6	0.102
53.0	60	13	0.217
56.9	62	18	0.290
60.8	56	28	0.500
64.8	63	52	0.825
68.7	59	53	0.898
72.6	62	61	0.984
76.5	60	60	1.000

At each dose x we can describe the number of beetles killed Y with a binomial distribution with probability parameter $p(x)$:

$$\mathbb{P}(Y = k) = \binom{n}{k} [p(x)]^k [1 - p(x)]^{n-k}$$

The Logistic Model

Since $p(x)$ is bounded between 0 and 1, we need a nonlinear link function to relate the mean $\mu = np(x)$ of the binomial distribution to the linear predictor. We use the logit function:

$$\log\left[\frac{\mu}{n-\mu}\right] \;=\; \log\left[\frac{p(x)}{1-p(x)}\right] \;=\; \beta_0 + \beta_1 x.$$

This is equivalent to an equation for the *odds* of a beetle getting killed:

$$\frac{p(x)}{1-p(x)} \;=\; \exp(\beta_0 + \beta_1 x),$$

and leads to the interpretation that increasing x by one unit changes the odds by a factor of e^{β_1}.

The coefficients β_0 and β_1 are found by maximizing the likelihood for this model:

$$\mathcal{L}(\beta_0, \beta_1) = \prod_{i=1}^{N} \binom{n_i}{k_i} \; [p(x_i)]^{k_i} \; [1-p(x_i)]^{n_i - k_i}$$

The Poisson Model

To illustrate the Poisson GLM we'll use a study of nesting horseshoe crabs. Each female crab had a male resident in her nest. The study investigated factors affecting the number of other males nearby ("satellites"). These factors included the female crab's color, spine condition, carapace width, and weight. (See Jupyter notebook.)

This is a counting problem, for which the Poisson distribution is appropriate:

$$\mathbb{P}(Y = n) \;=\; \frac{\mu^n}{n!}\, e^{-\mu}.$$

The mean of this distribution is μ and must be a positive number. Hence the loglinear link function is appropriate. For one factor x this is:

$$\log[\mu(x)] \;=\; \beta_0 + \beta_1\, x, \qquad \text{or} \qquad \mu(x) \;=\; \exp(\beta_0 + \beta_1\, x),$$

and we have the interpretation that increasing x by one unit multiplies the mean by a factor e^{β_1}. The coefficients β_0, β_1 are obtained by maximizing the likelihood function:

$$\mathcal{L}(\beta_0, \beta_1) \;=\; \prod_{i=1}^{N} \frac{[\mu(x_i)]^{n_i}}{n_i!}\, e^{-\mu(x_i)}$$

Regularization of Generalized Linear Models

In practice, the regression coefficients are found by maximizing the log of the likelihood rather than the likelihood itself. Thus in the logistic case:

$$\log \mathcal{L}(\beta_0, \beta_1, \ldots) = \sum_{i=1}^{N} \left\{ \log \binom{n_i}{k_i} + k_i \log \left[p(x_i) \right] + (n_i - k_i) \log \left[1 - p(x_i) \right] \right\}$$

and in the Poisson case:

$$\log \mathcal{L}(\beta_0, \beta_1, \ldots) \; = \; \sum_{i=1}^{N} \Big\{ n_i \log \left[\mu(x_i) \right] - \log \left[n_i! \right] - \mu(x_i) \Big\}$$

A regularization term can always be added to these log-likelihoods (Lasso, Ridge, or Elastic Net).

Testing Generalized Linear Models

- GLMs are estimated by the method of maximum likelihood. The likelihood also plays a role in testing. First define the deviance:

$$D \;=\; 2\;\left[\log \mathcal{L}_{\text{sat}} - \log \mathcal{L}_{\text{model}}\right],$$

  where $\mathcal{L}_{\text{model}}$ is the likelihood of the proposed model and $\mathcal{L}_{\text{sat}}$ is the likelihood of the *saturated* model, i.e. the model that predicts exactly the observations (it has as many parameters as observations). The deviance is a generalization of the residual sum of squares for the linear model. Asymptotically, it is distributed as a chisquared for $n - p - 1$ degrees of freedom, if the proposed model is a good approximation to the truth.
- Another possibility is to use a generalization of R^2:

$$R^2 = 1 - \frac{D}{D_0},$$

  where D_0 is the null deviance, that is the deviance for the model that only includes an intercept.

Example Deviance Calculation for the Poisson Case

Start with the likelihood for the proposed model:

$$\mathcal{L} \;=\; \prod_{i=1}^{N} \frac{[\mu(x_i)]^{n_i}}{n_i!}\, e^{-\mu(x_i)} \quad \text{and} \quad \log \mathcal{L} \;=\; \sum_{i=1}^{N} \left[n_i \log[\mu(x_i)] - \mu(x_i) - \log(n_i!) \right].$$

For the saturated model we introduce as many β_j coefficients in $\mu(x)$ as there are observations (N). Maximizing the likelihood then yields $\mu(x_i) = n_i$, so the saturated model likelihood is given by:

$$\mathcal{L}_{\text{sat}} \;=\; \prod_{i=1}^{N} \frac{n_i^{n_i}}{n_i!}\, e^{-n_i} \quad \text{and} \quad \log \mathcal{L}_{\text{sat}} \;=\; \sum_{i=1}^{N} \left[n_i \log(n_i) - n_i - \log(n_i!) \right]$$

The deviance is then:

$$D \;=\; 2(\log \mathcal{L}_{\text{sat}} - \log \mathcal{L}) \;=\; 2 \left[\sum_{i=1}^{N} \left(n_i \log \frac{n_i}{\mu(x_i)} + \mu(x_i) - n_i \right) \right]$$