Chapter 12

Generalized Linear Models

Generalized linear models (GLMs) are a class of supervised learning models that form a convenient bridge between machine learning and traditional statistics. The basic idea behind a GLM is that your outcome variable (a.k.a. response variable, see Chapter 2), y, follows a probability distribution. The expected value, or mean, of that distribution is related to the values of the predictors (a.k.a. covariates; see Chapters 2 and 3), $x_1, \dots, x_p$ in a model-specific way.

Linear and logistic regression, which we have already seen in Chapters 2, 3, 8, and 9, are both GLMs. Linear regression models data in which the outcome, y, is numeric ($y \in \mathbb{R}$) and follows a normal distribution. Logistic regression models data in which the outcome is binary ($y \in \{0, 1\}$) and follows a Bernoulli distribution[1] There are many more GLMs corresponding to outcomes that follow other types of probability distributions. For example, **loglinear (Poisson) regression** models data in which the outcome is a positive integer, or count ($y \in \{0, 1, 2, \dots\}$).

[1] In grouped logistic regression, the outcome follows a binomial distribution. More on that later.

12.1 Model Assumptions

In GLMs, the predictors can be anything – interval, ordinal, or nominal – regardless of the specific model one chooses. However, there are several other assumptions that are important to consider before fitting one of these models:

- We assume that the outcome follows a certain type of distribution (e.g. Bernoulli distribution for a logistic regression model, normal for linear, etc.) conditional on the predictors. This assumption is baked into the model structure. It is, therefore, important to consider whether the outcome distribution you chose actually makes sense for your particular problem. It is generally not advisable to use a linear regression model, for example, when your outcome is a count.
- We assume that the predictors are fixed and known, and thus have no error associated with their measurements[2].
- We assume that the predictors enter the model as a linear combination. This is why GLMs are referred to as "linear models".
- We assume that the n samples in our dataset are collected independently, so that the errors of the n sample outcomes are uncorrelated[3].

12.2 Modeling the Predictors

All of the GLMs we will see today incorporate a **linear combination** of predictors. A linear combination is an expression constructed from a set of terms by multiplying each term by a constant and adding the results. We denote the number of predictors in the model by p and the vector of predictors by x,

[2] Bayesian versions of these models relax this assumption, but we will not encounter these until much later

[3] Think back to our formulation of the likelihood in Chapter 5 and how it depended on the samples' being independent and identically distributed, or iid.

where

$$x = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

and we have included a "1" as the first element to allow for an **intercept**. We write $x^{(i)}$ to denote the vector of predictors associated with the ith training example. The coefficients of the linear combination (i.e. the model parameters we are hoping to learn) are denoted by:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$$

and we often express the linear combination as an inner product, written as

$$\beta^T x = \beta_0 + \sum_{j=1}^{p} \beta_j x_j.$$

Generalized linear models model the **expected value** of the outcome, $E[y]$, as a function of this linear combination of predictors. The function that relates the two is called the **link function**. Different types of GLM use different link functions.

12.3 Linear Regression

The linear regression model has a long history of development before the advent of GLMs, so it's typically taught in its own course with all of the associated model diagnostics, goodness of fit tests, etc. long before a student ever sees other GLMs. I think a comparative approach is more effective, which is why we're doing it this way[4].

12.3.1 Modeling the Outcome

In linear regression, we assume that the outcome, y, follows a normal distribution (see Section 4.2), conditional on the values of the predictors. Recall that the normal distribution is a continuous probability distribution with the following properties:

$$p(y|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \qquad E[y|\mu,\sigma] = \mu \qquad \text{var}(y|\mu,\sigma) = \sigma^2$$

where $y \in \mathbb{R}$.

12.3.2 Linking the Predictors to the Outcome

In linear regression, the mean of the outcome distribution, which is normal, can be any real number. We therefore use the **identity link**, setting $E[y]$ directly equal to the linear combination of predictors. Since the outcome is normal, we know that $E[y] = \mu$, the mean of the normal distribution. We therefore write:

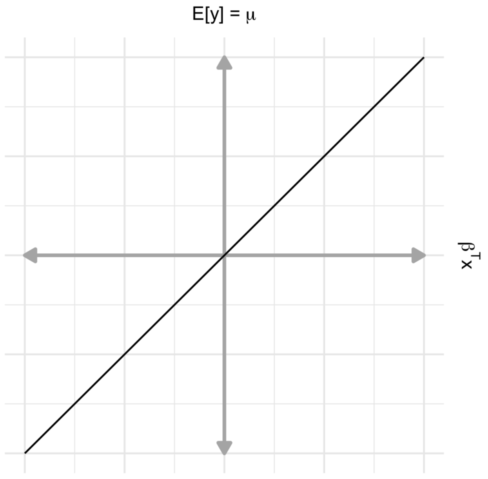
$$E[y] = \mu = \beta^T x \tag{12.1}$$

which is usually rearranged and rewritten as:

$$y = \beta^T x + \varepsilon$$

where $\varepsilon \sim N(0, \sigma^2)$. The relationship between $E[y]$ and $\beta^T x$ is shown below.

[4] The other thing about linear regression models is that they are usually fit using least squares methods instead of maximum likelihood. The parameter estimates are the same in both cases, as we will see much later.



12.4 Logistic Regression

Logistic regression models data where the outcome is binary; i.e. where y is "yes" or "no". Variants of logistic regression, called **multinomial logistic regression** and the **proportional odds model**, can also be used to model data where the outcome contains multiple categories that either have an ordering (ordinal) or do not (nominal). We will see how this works in a second.

12.4.1 Modeling the Outcome

In logistic regression the outcome, y, is either 0 or 1. We model it using the Bernoulli distribution (see Section 4.3), which is a discrete probability distribution with the following properties:

$$p(y|\mu) = \mu^y(1-\mu)^{1-y} \qquad E[y|\mu] = \mu \qquad \text{var}(y|\mu) = \mu(1-\mu)$$

where $y \in \{0, 1\}$.

12.4.2 Linking the Predictors to the Outcome

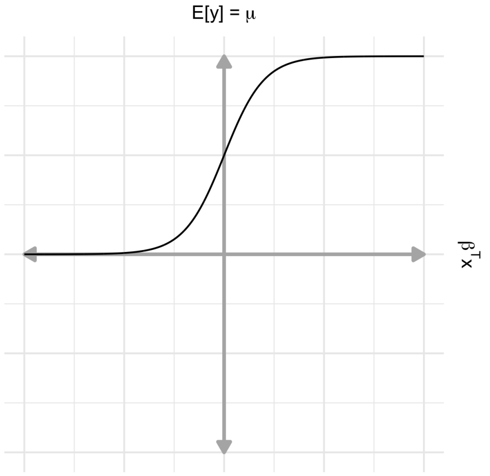
In logistic regression, the mean of the outcome distribution, which is Bernoulli, is a probability. It must therefore be a real number between 0 and 1. No matter how large or small $\beta^T x$ gets, the value of $E[y] = \mu$ cannot be outside this range. We therefore apply the **logistic function**, $f(x) = 1/(1 + \exp(-x))$, which has the range $(0, 1)$, to $\beta^T x$ to squash it:

$$E[y] = \mu = \frac{1}{1 + \exp\left(-\beta^T x\right)} \tag{12.2}$$

The relationship between $E[y]$ and $\beta^T x$ is shown below. We typically invert the model to write

$$\log \frac{\mu}{1-\mu} = \beta^T x$$

which is the standard form of the logistic regression model. The function $\log\left(\mu/(1-\mu)\right)$ is called the logit, and in logistic regression we say we use the **logit link**.



12.5 Poisson Regression

In Poisson regression, the outcome is a count. This type of regression is less common than linear and logistic regression, but we include it here mainly so you can see how the ideas from GLM extend to many different classes of outcome distributions within the exponential family.

12.5.1 Modeling the Outcome

In Poisson regression, we model the outcome using the Poisson distribution, which is a discrete probability distribution with the following properties:

$$p(y|\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!} \qquad E[y|\lambda] = \lambda \qquad \mathrm{var}(y|\lambda) = \lambda$$

where $y \in 0, 1, 2, \ldots$.

12.5.2 Linking the Predictors to the Outcome

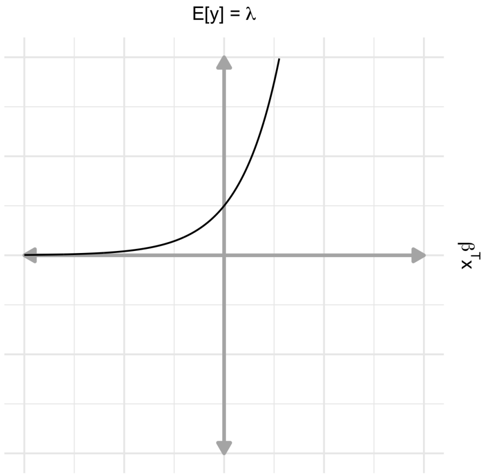
In Poisson regression, the mean of the outcome distribution, which is Poisson, is the expected value of a count. It must therefore be a real number greater than or equal to zero. In particular, no matter how small $\beta^T x$ gets, the value of $E[y] = \lambda$ cannot be negative. We therefore exponentiate $\beta^T x$ to ensure that the result is greater than zero:

$$E[y] = \lambda = \exp(\beta^T x) \tag{12.3}$$

The relationship between $E[y]$ and $\beta^T x$ is shown below. We typically invert the model to write

$$\log(\lambda) = \beta^T x$$

which is the standard form of the Poisson regression model. We say we use the **log link**.



12.6 Maximum Likelihood for GLMs

GLMs are typically fit using maximum likelihood estimation (see Chapter 5). A full treatment of MLE for GLMs is outside the scope of these notes, but I've put the start of the calculations for each type of model below.

12.6.1 Linear Regression

The likelihood for the linear regression model is:

$$\mathcal{L}(\mu^{(1)}, \ldots, \mu^{(n)}, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y^{(i)} - \mu^{(i)})^2}{2\sigma^2}\right]$$

where we use $\mu^{(i)}$ to represent the model's estimate of the mean of the outcome at the position of training example i. We can use Equation 12.1 to rewrite this as a function of the predictors:

$$\mathcal{L}(\beta, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y^{(i)} - \beta^T x^{(i)})^2}{2\sigma^2}\right]$$

Taking the log, we obtain the log-likelihood:

$$\log \mathcal{L}(\beta, \sigma) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y^{(i)} - \beta^T x^{(i)}\right)^2$$

Taking derivatives of the log-likelihood with respect to the βs, we find that we can maximize the likelihood by minimizing the sum-squares: $\sum_{i=1}^{n}\left(y^{(i)} - \beta^T x^{(i)}\right)^2$.

12.6.2 Logistic Regression

The likelihood for the logistic regression model is:

$$\mathcal{L}(\mu^{(1)}, \ldots, \mu^{(n)}) = \prod_{i=1}^{n} \mu^{(i)^{y^{(i)}}}(1 - \mu^{(i)})^{1-y^{(i)}}$$

Rewriting this as a function of the predictors, we get:

$$\mathcal{L}(\beta) = \prod_{i=1}^{n}\left(\frac{1}{1+\exp(-\beta^T x^{(i)})}\right)^{y^{(i)}}\left(\frac{\exp(-\beta^T x^{(i)})}{1+\exp(-\beta^T x^{(i)})}\right)^{1-y^{(i)}}$$

Taking the log, we obtain the log-likelihood:

$$\log \mathcal{L}(\beta) = \sum_{i=1}^{n}\left[-y^{(i)}\log\left[1+\exp(-\beta^T x^{(i)})\right] + (1-y^{(i)})\log\left[1+\exp(-\beta^T x^{(i)})\right]\right]$$

Again, we will take derivatives of the log-likelihood with respect to the βs to maximize it. However, we cannot solve for the optimal βs analytically; numerical optimization methods are used to perform the optimization.

12.6.3 Loglinear (Poisson) Regression

The likelihood for the Poisson regression model is:

$$\mathcal{L}(\lambda^{(1)}, \ldots, \lambda^{(n)}) = \prod_{i=1}^{n}\frac{\lambda^{(i)^{y^{(i)}}} e^{-\lambda^{(i)}}}{y^{(i)}!}$$

Rewriting this as a function of the predictors, we get:

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} \frac{\exp{(y^{(i)}\beta^T x^{(i)})}e^{-\exp{(\beta^T x^{(i)})}}}{y^{(i)}!}$$

Taking the log, we obtain the log-likelihood:

$$\log \mathcal{L}(\beta) = \sum_{i=1}^{n} \left[y^{(i)}\beta^T x^{(i)} - \exp(\beta^T x^{(i)}) - \log(y^{(i)}!) \right]$$

As with logistic regression, we cannot solve for the optimal βs analytically; numerical optimization methods are used.

12.7 Standard Errors and Hypothesis Tests

The magnitudes of the coefficients in these models matter only in relation to:

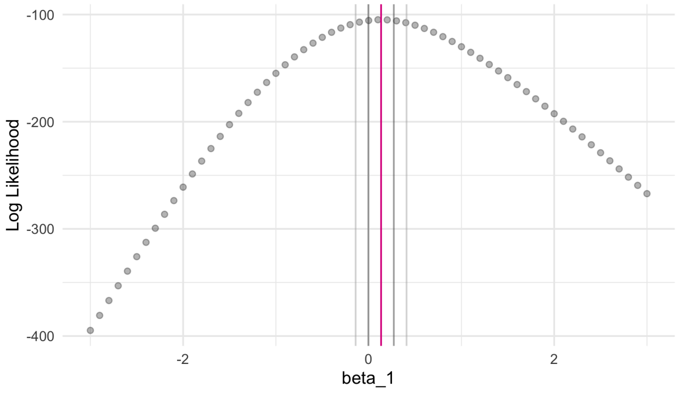
1. The scale on which the predictors are measured.
2. The amount of uncertainty the model has about their values.

For example, if a predictor varies only across a tiny range of values, its model coefficient may be large, since it quantifies the change in the link-function-transformed outcome when the predictor changes by 1.0. However, that doesn't mean that the predictor itself is important to the outcome[5].

Similarly, the model may be highly uncertain about a coefficient's value, owing to factors like a small dataset (small n) or collinearity among the predictors. Mathematically, high uncertainty means that the value of the likelihood doesn't change very rapidly as you move away from the maximum likelihood estimate of a coefficient. For example, here is how the log-likelihood for the logistic regression example above changes when we vary β_1 (the

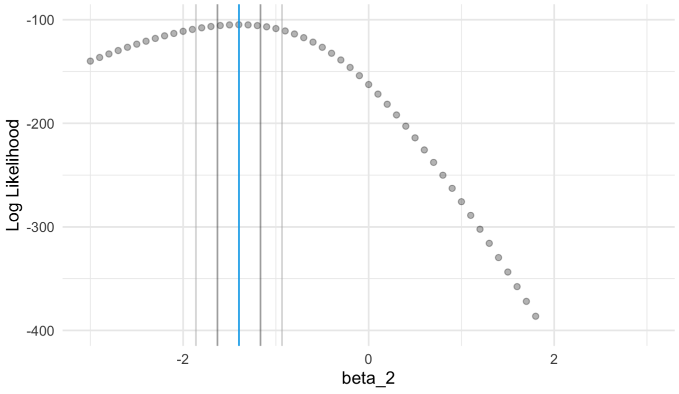
[5] This is one reason many advocate **scaling** and **centering** predictors before fitting a model. Centering means subtracting the mean value of a predictor from all of its individual measurements so that the mean of each centered predictor is zero. Scaling means dividing the values of each predictor by their standard deviation, so that the standard deviation of each predictor is 1.0. This enables the relative magnitudes of the model coefficients to be compared directly.

coefficient of x_1), keeping β_0 (the intercept) and β_2 (the coefficient of x_2) fixed at their MLEs:



The gray vertical lines are related to the **standard error** of the model coefficient, which is in turn related to the "flatness" of the likelihood surface around the MLE. The gray lines are situated at 1 and 2 standard errors away from the MLE in either direction. You can see that in the case of β_1, the gray lines overlap zero. The value zero (no effect) is a plausible estimate of the impact of x_1 on the outcome.

Contrast this with how the log-likelihood varies around the MLE for β_2:



Here the standard error is larger, but the magnitude of the coefficient is also larger, so the range of the gray lines does not overlap zero. These findings are reflected in the relative values of the **Z-statistic** (`z value`) and **P-value**

(`Pr(>z|)|`) in the model output for the two coefficients. Whether a coefficient's value is likely to be nonzero is typically evaluated using a formalism called a **hypothesis test**. We will discuss hypothesis tests in much greater detail in Chapter 6.

12.8 Example: Nesting Horseshoe Crabs Dataset

These data come from a study of nesting horseshoe crabs. Each of the 173 observed female horseshoe crabs had a male crab resident in her nest. The study investigated factors affecting whether the female crab had any other males, called *satellites*, residing nearby. (Source: Agresti, *Categorical Data Analysis*, Table 4.3. Data courtesy of Jane Brockmann, Zoology Department, University of Florida; study described in *Ethology* **102**: 1-21, 1996.)

```
SATELL  Number of satellites
COLOR   Color of the female crab
        (1 = light medium, 2 = medium, 3 = dark medium,
        4 = dark)
SPINE   Spine condition
        (1 = both good, 2 = one work or broken,
        3 = both worn or broken)
WIDTH   Carapace width of the female crab (cm)
WEIGHT  Weight of the female crab (g)
```

The GLM output of this model is:

```
Call:
glm(formula = satell ~ color + spine + width + weight, family = "poisson",
    data = d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0126  -1.8846  -0.5406   0.9448   4.9602

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.3435447  0.9684204  -0.355  0.72278
```

```
color       -0.1849325  0.0665236  -2.780  0.00544 **
spine        0.0399764  0.0568062   0.704  0.48160
width        0.0275251  0.0479425   0.574  0.56588
weight       0.0004725  0.0001649   2.865  0.00417 **
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 551.85  on 168  degrees of freedom
AIC: 917.15

Number of Fisher Scoring iterations: 6
```

Question 12.1

Comment on how the variables `color` and `spine` are coded here. Does this make sense in light of what those variables mean?

Question 12.2

Interpret the values of each of these coefficients. Based on the coefficient values and their standard errors, which predictor(s) do you think have the greatest impact on the number of male satellites around a nesting female horseshoe crab?