Chapter 12

Generalized Linear Models

Linear and logistic regression, which we have seen already in Chapters 2, 3, 8, and 9, are members of a broader class of supervised learning models called **generalized linear models (GLMs)**. In GLMs, the outcome variable, y, is assumed to follow a probability distribution of a particular type. For example, in linear regression, y follows a normal distribution. In logistic regression, y is binary ($y \in \{0, 1\}$) and follows a Bernoulli distribution[1]. The expected value, or mean, of the outcome distribution is related to a **linear combination** of the predictors, $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$, via a model-specific **link function**.

GLMs, like maximum likelihood (Chapter 5), are normally considered an advanced topic. However, they provide a nice example of how the same formalism – modeling the response variable using a probability distribution, assuming a certain form for the predictors, optimizing the whole thing using maximum likelihood – can be applied to solve different-looking problems. They are also a good entryway into the sorts of optimization tasks performed by graphical models and deep learning algorithms.

12.1 Model Assumptions

GLMs require us to make several assumptions which affect both our choice of model and our interpretation of model output:

[1] In **grouped** logistic regression, it follows a binomial distribution.

1. 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.

2. We assume that the predictors are fixed and known, and thus have no error associated with their measurements[2].

3. We assume that the predictors enter the model as a linear combination, $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$. This is why GLMs are referred to as "linear models".

4. 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 Notation for the Predictors

As mentioned above, GLMs assume that the predictors enter the model as a linear combination. 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, 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

[2]Bayesian versions of these models relax this assumption.

[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.

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**, or **dot product**, of the two vectors, written as

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

This is just notational shorthand.

Question 12.1

We saw the details of linear and logistic regression models in Chapters 8 and 9 and discussed the limitations of predictors' entering as a linear combination. What are some of those limitations?

Question 12.2

Just to confirm that you understand this notation, write out the form of $\beta^T x$ for a model with (a) one predictor, (b) three predictors. Write both the general form and the form for one training example, $x^{(i)}$.

12.3 Modeling the Outcome

Generalized linear models model the expected value of the outcome, $E[y]$, as a function of this linear combination of predictors.

12.3.1 Linear Regression

In linear regression, we assume that the outcome, y, follows a normal distribution (see Section 4.2), whose mean is controlled by the values of the predictors.

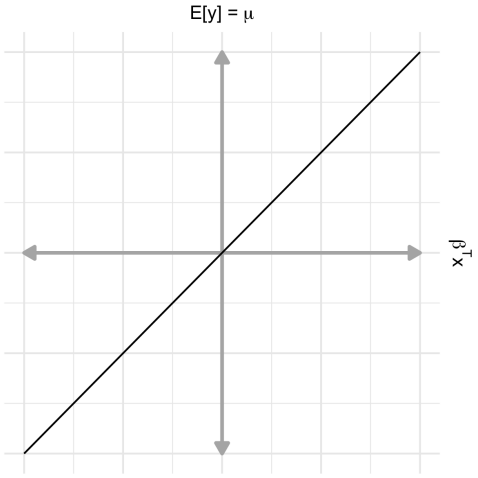
Recall that the normal distribution is a continuous probability distribution with the following properties:

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

where $y \in \mathbb{R}$. Its mean, μ, can be any real number. To link μ to the predictors, therefore, we simply set it equal to $\beta^T x$, like so:

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

This is called using the **identity link**. The relationship between $E[y] = \mu$ and $\beta^T x$ is shown below.



Question 12.3

In this model, how much does the mean of the outcome distribution, μ, change as you vary each predictor? For example, if you have $p = 3$ predictors, by how much does μ change as the value of x_2 changes by one unit (for example, from 1 to 2)? How much does μ change as the value of x_2 changes from 3 to 4? What about x_1 and x_3?

12.3.2 Logistic Regression

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:

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

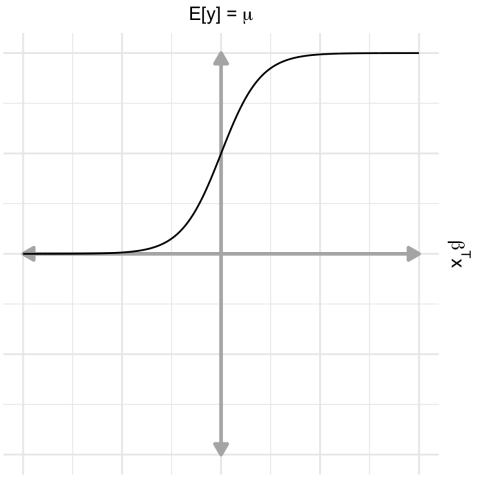
where $y \in \{0, 1\}$. Because μ is a probability, it must be a real number between 0 and 1. No matter how large or small $\beta^T x$ gets, the value of $E[y] = \mu$ cannot be outside this range. We therefore apply the **logistic function**, $f(x) = 1/(1+\exp(-x))$, which has the range $(0, 1)$, to $\beta^T x$ 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

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

The function $\log\left(\mu/(1-\mu)\right)$ is called the logit, and we say we use the **logit link**.



Question 12.4

Let's revisit Question 9.1. Now that we've described logistic regression in the framework of GLMs, what more can you say about why the model is not of the form

$$\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p?$$

12.3.3 Poisson Regression

In Poisson regression, the outcome is a count. We model the outcome using a Poisson distribution, which is a discrete probability distribution with the following properties (Section 4.5):

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

where $y \in 0, 1, 2, \ldots$. Because λ, the mean of the outcome distribution, is the expected value of a count, it must 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 λ 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

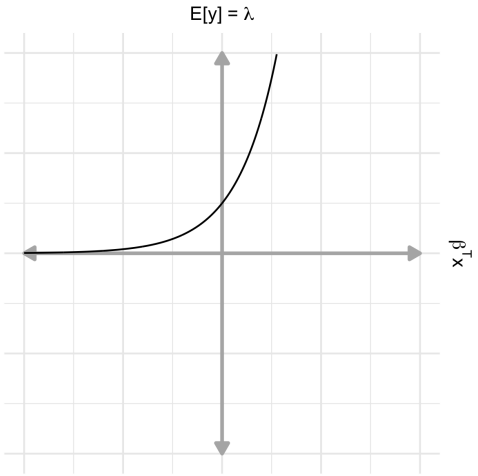
$$\log(\lambda) = \beta^T x$$

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

Question 12.5

There are many other generalized linear models. In each case, the mean (expected value) of a probability distribution is related, via a link function, to a linear combination of the predictors.

Knowing this, how would you create a GLM where the outcome follows an exponential distribution (Section 4.7)? Which link would you use?



12.4 Maximum Likelihood for GLMs

GLMs are 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. The only difference between these calculations and those in Chapter 5 is that now our parameters of interest, the means of our outcome distributions, are functions of our predictors $x_1, \ldots, x_p$. Our job is to find the coefficients on those predictors, $\beta_0, \ldots, \beta_p$, that provide the best fit between our model and our training data.

12.4.1 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: $\sum_{i=1}^{n}\left(y^{(i)} - \beta^T x^{(i)}\right)^2$.

Question 12.6

Take a minute to stare at this result. When most people learn linear regression, they learn that these models are fitted by minimizing the sum of squared residuals (see Chapter 8). Indeed, linear regression models predate GLMs and are typically fit using ordinary least squares, not maximum likelihood. If you fit a linear regression model in R using the `lm` package, you're using OLS. If you use the `glm` package with the argument `family = "gaussian"`, you're using maximum likelihood. However, both methods will produce the same fitted models. Do you see why this is?

12.4.2 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 in this case. Numerical optimization methods are used to find the maximum likelihood estimates, $\hat{\beta}_0, \hat{\beta}_1$, etc.

12.4.3 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:

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

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.

Question 12.7

Think of the log-likelihood as measuring the height of a hill. Your data,

$$\left\{ x^{(1)}, \ldots, x^{(n)} \right\}$$

don't change, so we don't care about their effect on the height. What we care about are the parameters, $\beta_0, \ldots, \beta_p$. For each combination of those $p + 1$ parameters, the height changes. We want to find the combination of parameters

that puts us at the top of the hill.

The first derivative of the log-likelihood with respect to one of the parameters, β_j, is

$$\frac{\partial \log \mathcal{L}}{\partial \beta_j}$$

and the vector of all of these first derivatives for $\beta_0, \ldots, \beta_j$ is called the **gradient**. Evaluated at a particular set of parameters, the gradient tells you how steep your hill is in the direction of each of your $p + 1$ parameters. How could you use this information to maximize the likelihood? You don't need to do any math. Just say how you would do it.

Question 12.8

There are many different numerical optimization algorithms that one can use to maximize the likelihood (i.e., find the top of the hill). One of them is called **Fisher scoring**. Examine the output of the logistic regression models in Chapter 9 and the Poisson regression model shown below in Section 12.6. Where do you see the term "Fisher scoring"? What do you think the term "Fisher scoring iterations" refers to?

12.5 Standard Errors and Hypothesis Tests

Here, once again, is the summary output from a logistic regression model of the ER readmissions example from Chapter 2, reprinted again in Section 9.1:

```
Call:
glm(formula = y ~ x1 + x2, family = "binomial", data = df)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.88232  -0.90614  -0.05965   0.86579   2.28489

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.9780     0.2945   3.321 0.000897 ***
x1            0.1344     0.1372   0.980 0.327272
x2           -1.3981     0.2316  -6.035 1.59e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 277.26  on 199  degrees of freedom
Residual deviance: 209.54  on 197  degrees of freedom
AIC: 215.54

Number of Fisher Scoring iterations: 4
```

As we discussed in Chapter 9, the magnitudes of the coefficients in these models matter, but they are only important in relation to:

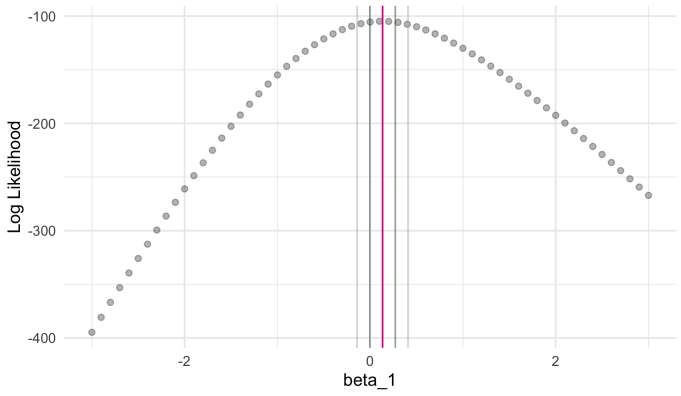
1. The scale on which the predictors are measured.
2. The amount of uncertainty the model has about their values.

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[4].

Similarly, the model may be highly uncertain about a coefficient's value, owing to factors like a small dataset (small n) or collinearity (correlations) 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

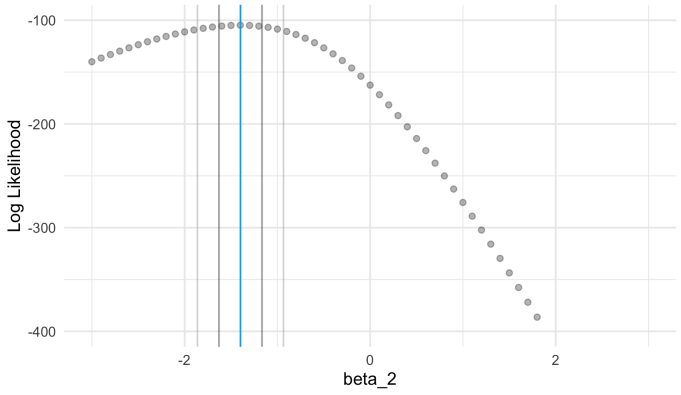
[4]This is one reason many advocate **scaling** and **centering** predictors before fitting a model. Centering means subtracting the mean value of a predictor from all of its individual measurements so that the mean of each centered predictor is zero. Scaling means dividing the values of each predictor 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.

log-likelihood for the logistic regression example above changes when we vary β_1 (the 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.

Question 12.9

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. With that in mind, let's reconsider Question 9.6. How do these likelihood plots and the null distributions shown in Question 9.6 convey the same information?

12.6 Example: Nesting Horseshoe Crabs Dataset

Let's examine some output from a Poisson regression model, which is a type of GLM with which you may not already be familiar.

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 12.10

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

Question 12.11

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?

Question 12.12

How could you use a decision tree to model the horseshoe crabs data? What are its advantages and disadvantages relative to Poisson regression (a type of GLM)?