An Introduction to Generalized Linear Models

Emma Grossman, Leah Marcus, Emily Palmer, Katherine Pulham, Andrew Rumments

2021-03-10

Contents

Chapter 1

Description of our book

Someone please update this.

We used (**?**)

(also cite data we use...)

Chapter 2

Introduction

2.1 What came before - Linear models

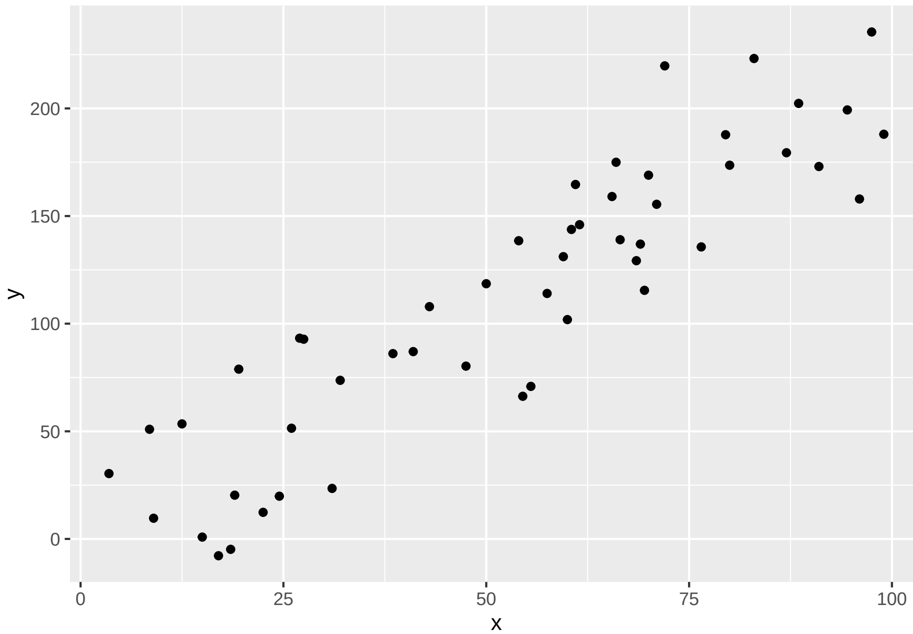
If you are reading this book, you might already be familiar with linear models. Given our data, if we make some key assumptions (see **??**), we can perform inference or prediction by assuming that our response value forms a linear relationship with our explanatory variable (or variables).

The reasoning of linear models is often intuitive, if we make a scatterplot our data, and see this:

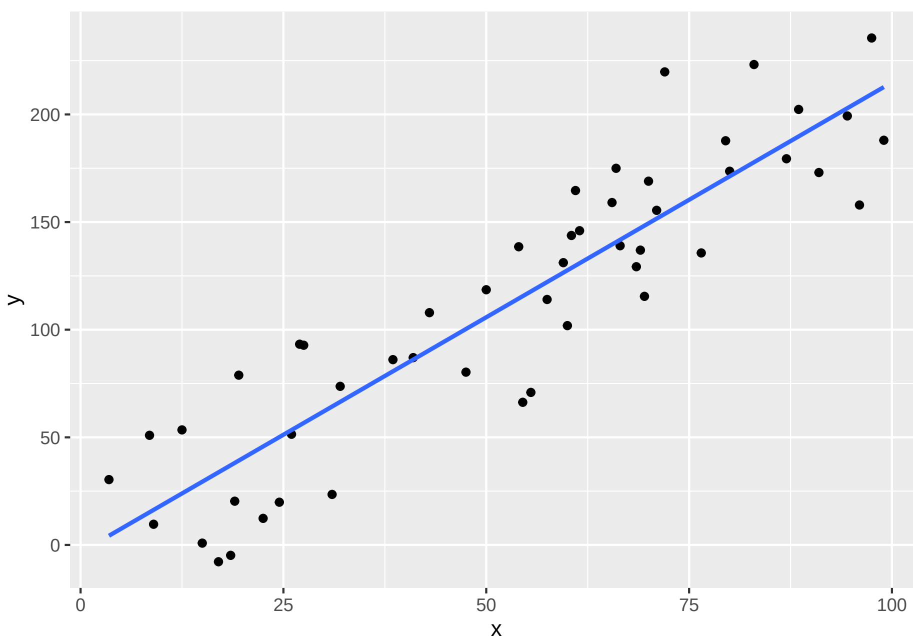
```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```



we might want to fit a straight line through the cloud of points, i.e. modeling the relationship linearly.



To interpret this relationship and make predictions, we need to know the slope and intercept of this line. This is done by minimizing the least squares, which will be explored in chapter 3 @ref{linear}.

2.2 Some definitions

Predictor - the thing on the y-axis Explanatory variable - the stuff on the x-axis. Note that we can have more than one (but won't plot it then), and then this becomes multivariate regression.

Something that is an estimated quantity will have a hat over it. For example, we might assume that there is some 'true' (but unknown) linear relationship between our explanatory variables and our predictor.

$$y = \beta_0 + \beta_1 x$$

From our sample data, we use a linear model to make an estimate of β_0 and β_1, so our estimate/best guess of this true model relationship is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

We of course want our $\hat{\beta}_0$ and $\hat{\beta}_1$ to be a 'good' and 'close' estimate of the unknown quantities β_0 and β_1. Ideas of what 'good' and 'close' mean will be covered in the next section.

2.3 Assumptions of linear models

A linear model might very well be a good model if our data look like **??**. However, there are many cases where it might be inappropriate to use a linear model. To understand these cases, we first review the assumptions of linear models.

Linear models assume:

- The relationship between the explanatory variables and the response is linear
- The samples are independent.
- The errors are normally distributed with mean 0 and constant variance

We can write these assumptions down in notation as such.

$$y_i = \beta_0 + \beta_1 x + \epsilon_i$$

where

$$\epsilon_i \sim \text{iid } N(0, \sigma^2)$$

In words, this means that each this means that the errors are independent and identically distributed by the normal distribution, with mean 0 and constant variance σ^2 (notice how there is no subscript i for the variance)

If these assumptions hold, we then write our model as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

How can we tell when these assumptions are violated?

- Knowledge of the data.
- Plots

2.4 What happens when we break the assumptions of linear models

Linear models are generally robust, and can be reasonable when assumptions are not exactly met. However, if we know assumptions are not met, and how they are not met, it is appropriate to use a more appropriate model for the data.

2.5 Random and Systematic Component

We will now analyze the assumptions for linear models and explore how we can generalize them. (and create generalized linear models!)

$$y_i = \beta_0 + \beta_1 x + \epsilon_i$$

$$\epsilon_i \sim \text{iid } N(0, \sigma^2)$$

We refer to the first equation as the Systematic Component, and the second equation as the Random Component.

A Generalized Regression Model has a systematic component:

$$g(y_i) = \beta_0 + \beta_1 x + \epsilon_i$$

To generalized the systematic component, we use a link function $g(y)$, so we now require some function of the response to be linearly related to our explanatory variables.

and a random component:

$$\epsilon_i \sim \text{ iid } EDM(\phi)$$

In words, the errors are independently distributed according to some probability distribution in the Exponential Dispersion Family, which will be discussed in the next chapter. Normal, Binomial, and Poisson distributions all fall into this family.

We note that normal linear models fall exactly into this framework, where $g(y_i) = y_i$ the identity function, and use the Normal distribution as our random component.

Deciding on what Random and Systematic component to use requires u

2.6 Random and Systematic components for Binary and Count data

The two most common cases of GRMs are those for Binary and Count data

For Binary data, the systemtatic component is, and the random component is: We call these types of GRMS logistic regression or …

For Count data, the systematic component is, and the random component is. We call these types of GRMS

2.7 Parameter estimation

The last difference between linear models and generalized linear models is the way we estimate the parameters β.

2.8 Conclusion

Linear models are not always the best tool for describing relationship in data. Luckily we can generalize the ideas and framework developed in linear models to hold for more general cases to create GLMs. Using a more general framework and more general assumptions allows us to build tools that will hold for all GRMs. The most notable of these that we will further explore are GRMs for binary data (ch4) and count data (ch5)

2.9 Examples

Perhaps some examples of data and students can tell what type of data it should be modeled by?

Chapter 3

How are GLMs "different"?

3.1 Introdution

So, we've talked about the issues that linear models can run into. The question now is how do we deal with these issues? What we're going to need to do is expand the type of model we're trying to fit. In linear regression we assumed two things: that the response variable Y_i is distributed normally, with constant variance σ^2, and that the mean of the response variable is a linear combination of the explanatory variables. These two assumptions can be stated as

1. $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$
2. $\mu_i = \beta_0 + \beta_1 X_{i,1} + ... + \beta_k X_{i,k}$

In this chapter we're going to make our model more general by expanding these two assumptions. The first assumption, which we will call the random component, is going to change from Y being distributed normally to Y being distributed according to *some probability family*. The second assumption is going to change from μ_i directly equaling the linear predictor to *some function* of μ_i being equal to this linear predictor.

3.2 Assumptions of a GLM

GLMs are made up of two components: a random component, and a structural component. In general, what we're saying is that the response variable of interest is a random variable that follows a specific probability distribution (random component). This probability distribution is, in some way, related to a linear combination of the explanatory variables (systematic component). This linear combination of the explanatory variables is where the "linear" in "generalized linear model" comes from. In linear regression, which is a special case of the generalized linear model, the random component is that Y comes from a normal distribution: $Y_i \sim N(\mu_i, \sigma^2)$ and the systematic component is that the mean is some linear combination of the explanatory variables: $\mu_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki}$. With GLMs, our goal is to extend this framework so that we're not just limited to the normal distribution for the random component of our model, for reasons we discussed in the last chapter.

However, when we fit these models, we need to be sure of a couple of things. We need to ensure that for a linear combination of explanatory variables, we can identify which distribution the response variable comes from. We also need to ensure that the parameters of that distribution we're trying to fit are estimable. To ensure that we're able to properly fit these models, GLMs consider a specific kind of family of distributions for the random component: the Exponential Dispersion Model.

3.3 Framework

3.3.1 Exponential Dispersion models

An exponential dispersion model is a specific type of random variable, whose pdf follows a specific form:

$$f_Y(y) = a(y, \phi) exp \left[\frac{y\theta - \kappa(\theta)}{\phi} \right]$$

in this form, θ is called the *canonical parameter*, and ϕ is called the *dispersion parameter*. For our purposes, the function $a(y, \phi)$ is not of much interest, but it is needed to guarantee that $f_Y(y)$ integrates to 1, and is therefore a valid probability density function. $\kappa(\theta)$ is called the *cumulant* function, and will be useful to us in estimation. Another term for an exponential dispersion model is to say that the family of random variables is an exponential family.

A surprising, and fortunate, number of families of distributions are exponential dispersion models. Notably, some of them are

- Normal random variables
- Bernoulli random variables
- Binomial random variables
- Poisson random variables
- Exponential random variables
- Gamma random variables
- Negative binomial random variables

We'll spare the details for most of these families, but to show the general idea for how we decide whether or not a family of random variables is an exponential dispersion model, we shall consider the poisson random variable.

Example: For a poisson random variable, the pmf is written as

$$f_Y(y) = e^{-\lambda}\frac{\lambda^y}{y!}$$

by applying the identity $x = e^{log(x)}$ to the numerator, we see that this is equivalent to

$$f_Y(y) = \frac{1}{y!}exp\left[-\lambda + ylog(\lambda)\right] = \frac{1}{y!}exp\left[\frac{ylog(\lambda) - \lambda}{1}\right]$$

and we see that the poisson random variable is an exponential dispersion model with dispersion parameter $\phi = 1$, with canonical parameter $\theta = log(\lambda)$ and with cumulant function $\kappa(\theta) = \lambda = e^\theta$. Notice how we left out the $\frac{1}{y!}$ term of the exponential because it was not needed to put the function into this important form. □

3.3.2 Properties of EDMs

Once we can get a probability distribution function into the exponential dispersion model form, we can connect this form to both the mean and variance of the random variable. The expected value (mean) of the random variable is simply the first derivative of the cumulant function with respect to the canonical parameter:

$$E[Y] = \mu = \frac{d}{d\theta}\kappa(\theta)$$

The cumulant function is also related to the variance of the random variable. The variance of the random variable is the dispersion parameter multiplied by the second derivative of the cumulant function with respect to the canonical parameter:

$$var(Y) = \phi\frac{d^2}{d\theta^2}\kappa(\theta)$$

The second part of this expression is an important quantity, called the variance function. Notice that it is equal to the first derivative of the expected value of Y as well:

$$V(\mu) = \frac{d^2}{d\theta^2}\kappa(\theta) = \frac{d}{d\theta}\mu$$

It is worth noting that, in addition to helping us estimate properties of Y, the variance function uniquely determines the family of distributions (type of random variable) for a given EDM. For instance, following our previous example, since $\kappa(theta) = e^\theta$, the variance function is $V(\mu) = \frac{d^2}{d\theta^2}e^\theta = e^\theta = \lambda = \mu$. What this means is that *any* EDM with variance function $V(\mu) = \mu$ will be a poisson random variable.

3.3.3 Linking the EDM to the explanatory data

Recall, just for a second, the goal of constructing these models. We have a response variable, Y_i, and a collection of explanatory variables $X_1, X_2, X_3, ...X_k$. We want to be able to look at a combination of the explanatory variables and draw some conclusions about Y. Perhaps we want to predict Y with a point estimator. If we make this sort of prediction, it's also of interest to know how precise that estimate will be, so we may wish to find an interval estimate for the prediction as well. Ultimately, all of these things come from the distribution of Y, so the thing that is of interest is to be able to know what the probability distribution of Y is given the input values of the X_i's.

As stated before, the L in GLM stands for linear, and these explanatory variables are where that linearity comes into play. In GLMs, we're assuming that the quantity we'll use to predict the response variable Y is a linear combination of the explanatory data $\beta_0 + \beta_1 X_1 + ... + \beta_k X_k$. We will call this quantity the linear predictor; a common shorthand way of writing it is to use the greek letter $\eta = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$. In practice, we often have multiple repetitions of the explanatory variables, where Y_i is a random variable who's distribution is somehow linked to the covariates $X_{1i}, X_{2i}, X_{3i}, ...X_{ki}$. In this case, we will denote the separate linear predictors as $\eta_i = X_{1i}, X_{2i}, X_{3i}, ...X_{ki}$. Note that although the variables may change, the coefficients $\beta_0,\ \beta_1,\ ...\ \beta_k$ are the same for every η_i. These β coefficients are the thing we must estimate to fit our GLM.

The question remains of *how* we connect η to the distribution of Y. First, we have to suppose what kind of distribution Y is coming from (is it a poisson random variable? Binomial?) and then we need to find some function g() such that the expected value $E[Y] = \mu$ is simply $g(\mu) = \eta$. For this, we have to place a couple restrictions on g. First, g must be a strictly monotonic function (strictly increasing or strictly decreasing) from some subset of the real numbers onto the set of all values that μ could be. We require the monotonicity to ensure that we don't have multiple separate means being linked to the same linear predictor. This function also has to be differentiable to make sure that the tools we use to estimate μ don't break. In practice, these requirements don't come up very much, since typically there are a couple of link functions that get used for each family of probability densities.

One special link function for each EDM family is the *canonical link function*. For an EDM family of distributions, the canonical link function is the function $g(\mu)$ that satisfies $\eta = \theta = g(\mu)$.

The canonical link function isn't the only valid link function. Take for example the binomial family of distributions, and let $Y \sim Binom(n, p)$, for some known n. Note that $\mu = p$. In this case, the set of possible values of p is the unit interval $(0, 1)$. The canonical link function for this family is the logit function:

$$g(p) = log\left(\frac{p}{1-p}\right)$$