Julie Legler and Paul Roback

Broadening Your Statistical Horizons

Generalized Linear Models and Multilevel Models

Contents

List of Tables

List of Figures

Preface

Placeholder

0

Review of Multiple Linear Regression

Placeholder

0.1 Learning Objectives

0.2 Introduction to Broadening Your Statistical Horizons

0.3 Ordinary Least Squares (OLS) Assumptions

0.3.1 Cases that do not violate the OLS assumptions for inference

0.3.2 Cases where the OLS assumptions for inference are violated

0.4 Review of Multiple Linear Regression

0.4.1 Case Study: Kentucky Derby

0.5 Initial Exploratory Analyses

0.5.1 Data Organization

0.5.2 Univariate Summaries

0.5.3 Bivariate Summaries

0.6 Multiple linear regression modeling

0.6.1 Simple linear regression with a continuous predictor

0.6.2 linear regression with a binary predictor

0.6.3 Multiple linear regression with two predictors

0.6.4 Inference in multiple linear regression: normal theory

0.6.5 Inference in multiple linear regression: bootstrapping

0.6.6 Multiple linear regression with an interaction term

0.6.7 Building a multiple linear regression model

0

Beyond Least Squares: Using Likelihoods to Fit and Compare Models

Placeholder

0.9 Learning Objectives

0.10 Case Study: Does sex run in families?

0.10.1 Research Questions

0.11 Model 0: Sex Unconditional Model (Equal probabilities, Independence)

0.12 Model 1: Sex Unconditional Model (Any Probability, Independence)

0.12.1 What is a likelihood?

0.12.2 Finding MLEs

0.12.2.1 Graphically approximating an MLE

0.12.2.2 Numerically approximating an MLE

0.12.2.3 MLEs using calculus (Optional)

0.12.2.4 How does sample size affect the likelihood?

0.12.3 Summary

0.12.4 Is a likelihood a probability function? (Optional)

0.13 Model 2: Sex Conditional Model (Sex Bias)

0.13.1 Model Specification

0.13.2 Application to Hypothetical Data

0.14 Case Study: Analysis of the NLSY data

0.14.1 Model Building Plan

0.14.2 Family Composition of Boys and Girls, NLSY: Exploratory Data Analysis

0.14.3 Likelihood for the Sex Unconditional Model: the NLSY data

0

Poisson Regression

Placeholder

0.27 Learning Objectives

0.28 Introduction to Poisson Regression

0.28.1 Poisson Regression Assumptions

0.28.2 A Graphical Look at Poisson Regression

0.29 Case Studies Overview

0.30 Case Study: Household Size in the Philippines

0.30.1 Data Organization

0.30.2 Exploratory Data Analyses

0.30.3 Estimation and Inference

0.30.4 Using Deviances to Compare Models

0.30.5 Using Likelihoods to fit Poisson Regression Models (Optional)

0.30.6 Second Order Model

0.30.7 Adding a covariate

0.30.8 Residuals for Poisson Models (Optional)

0.30.9 Goodness-of-fit

0.31 Least Squares Regression vs. Poisson Regression

0.32 Case Study: Campus Crime

0.32.1 Data Organization

0.32.2 Exploratory Data Analysis

0.32.3 Accounting for Enrollment

0

Generalized Linear Models (GLMs): A Unifying Theory

0.38 Learning Objectives

- Determine if a probability distribution can be expressed in one-parameter exponential family form.
- Identify canonical links for distributions of one parameter exponential family form.

0.39 One parameter exponential families

Thus far, we have expanded our repertoire of models from OLS to include Poisson regression. But in the early 1970s Nelder and Wedderburn [1972] identified a broader class of models that generalizes the multiple linear regression we considered in the introductory chapter and are referred to as **generalized linear models (GLMs)**. All GLMs have similar forms for their likelihoods, MLEs, and variances. This makes it easier to find model estimates and their corresponding uncertainty. To determine whether a model is a GLM, we consider the following properties. When a probability formula can be written in the form below

$$f(y;\theta) = e^{[a(y)b(\theta)+c(\theta)+d(y)]} \tag{0.1}$$

and if the **support** (the set of possible input values) does not depend upon θ, it is said to have a **one-parameter exponential family form**. We demonstrate that the Poisson distribution is a member of the one parameter exponential family by writing its probability mass function (pmf) in the form of Equation (0.1) and assessing its support.

0.39.1 One Parameter Exponential Family: Possion

Recall we begin with

$$P(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!} \quad \text{where} \quad y = 0, 1, 2 \ldots \infty$$

and consider the following useful identities for establishing exponential form:

$$a = e^{\log(a)}$$

$$a^x = e^{x\log(a)}$$

$$\log(ab) = \log(a) + \log(b)$$

$$\log\left(\frac{a}{b}\right) = \log(a) - \log(b)$$

Determining whether the Poisson model is a member of the one-parameter exponential family is a matter of writing the Poisson pmf in the form of Equation (0.1) and checking that the support does not depend upon λ. First, consider the condition concerning the support of the distribution. The set of possible values for any Poisson random variable is $y = 0, 1, 2 \ldots \infty$ which does not depend on λ. The support condition is met. Now we see if we can rewrite the probability mass function in one-parameter exponential family form.

$$\begin{aligned} P(Y = y) &= e^{-\lambda}e^{y\log\lambda}e^{-\log(y!)} \\ &= e^{y\log\lambda - \lambda - \log(y!)} \end{aligned} \tag{0.2}$$

The first term in the exponent for Equation (0.1) must be the product of two factors, one solely a function of y, $a(y)$, and another, $b(\lambda)$, a function of λ only. The middle term in the exponent must be a function of λ only; no $y's$ should appear. The last term has only $y's$ and no λ. Since this appears to be the case here, we can identify the different functions in this form:

$$a(y) = y \tag{0.3}$$

$$b(\lambda) = \log(\lambda) \tag{0.4}$$

$$c(\lambda) = -\lambda$$

$$d(y) = -\log(y!) \tag{0.5}$$

$$\tag{0.6}$$

These functions have useful interpretations in statistical theory. We won't be going into this in detail, but we will note that function $b(\lambda)$, or more generally $b(\theta)$, will be particularly helpful in GLMs. The function $b(\theta)$ is referred to as the **canonical link**. The canonical link is often a good choice to model as a linear function of the explanatory variables. That is, Poisson regression should be set up as $\log(\lambda) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$. In fact, there is a distinct advantage

to modeling the canonical link as opposed to other functions of θ, but it is also worth noting that other choices are possible, and at times preferred, depending upon the context of the application.

There are other benefits of identifying a response as being from a one parameter exponential family. For example, by creating an unifying theory for regression modeling, Nelder and Wedderburn made possible a common and efficient method for finding estimates of model parameters using iteratively reweighted least squares (IWLS). In addition, we can use the one parameter exponential family form to determine the expected value and standard deviation of Y. With statistical theory you can show that

$$\mathrm{E}(Y) = -\frac{c'(\theta)}{b'(\theta)} \quad \text{and} \quad \mathrm{Var}(Y) = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3}$$

where differentiation is with respect to θ. Verifying these results for the Poisson response:

$$\mathrm{E}(Y) = -\frac{-1}{1/\lambda} = \lambda \quad \text{and} \quad \mathrm{Var}(Y) = \frac{1/\lambda^2}{(1/\lambda^3)} = \lambda$$

We'll find that other distributions are members of the one parameter exponential family by writing their pdf or pmf in this manner and verifying the support condition. For example, we'll see that the binomial distribution meets these conditions, so it is also a member of the one parameter exponential family. The normal distribution is a special case where we have two parameters, a mean μ and standard deviation σ. If we assume, however, that one of the parameters is known, then we can show that a normal random variable is also from a one parameter exponential family.

0.39.2 One parameter exponential family: Normal

Here we determine whether a normal distribution is a one parameter exponential family member. First we will need to assume that σ is known. Next, possible values for a normal random variable range from $-\infty$ to ∞, so the support does not depend on μ. Finally, we'll need to write the probability density function (pdf) in the one parameter exponential family form. We start with the familiar form:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/(2\sigma^2)}$$

Even writing $1/\sqrt{2\pi\sigma^2}$ as $e^{-\log\sigma - \log(2\pi)/2}$ we still do not have the pdf written in one parameter exponential family form. We will first need to expand the exponent so that we have

$$f(y) = e^{[-\log\sigma - \log(2\pi)/2]} e^{[-(y^2 - 2y\mu + \mu^2)/(2\sigma^2)]}$$

Without loss of generality, we can assume $\sigma = 1$, so that

$$f(y) \propto e^{y\mu - \frac{1}{2}\mu^2 - \frac{1}{2}y^2}$$

and $a(y) = y$, $b(\mu) = \mu$, $c(\mu) = -\frac{1}{2}\mu^2$, and $d(y) = -\frac{1}{2}y^2$.

From this result, we can see that the canonical link for a normal response is μ which is consistent with what we've been doing with OLS, since the simple linear regression model has the form:

$$\mu_{Y|X} = \beta_0 + \beta_1 X.$$

0.40 Generalized Linear Modeling

GLM theory suggests that the canonical link can be modeled as a linear combination of the explanatory variable(s). This approach unifies a number of modeling results used throughout the text. For example, likelihoods can be used to compare models in the same way for any member of the one-parameter exponential family.

We have now **generalized** our modeling to handle non-normal responses. In addition to normally distributed responses, we are able to handle Poisson responses, binomial responses, and more. Writing a pmf or pdf for a response in one parameter exponential family form reveals the canonical link which can be modeled as a linear function of the predictors. This linear function of the predictors is the last piece of the puzzle for performing generalized linear modeling. But, in fact, it is really nothing new. We already use linear combinations and the canonical link when modeling normally distributed data.

Three Components of a GLM

1. Distribution of Y (e.g., Poisson)
2. Link Function (a function of the parameter, e.g., $\log(\lambda)$ for Poisson)
3. Linear Predictor (choice of predictors, e.g., $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$)

TABLE 0.1: One parameter exponential family form and canonical links.

Distribution	One-parameter Exponential Family Form	Canonical Link
Binary		
Binomial		$\text{logit}(p)$
Poisson	$P(Y=y) = e^{y \log \lambda - \lambda - y!}$	$\log(\lambda)$
Normal	$f(y) \propto e^{y\mu - \frac{1}{2}\mu^2 - \frac{1}{2}y^2}$	μ
Exponential		
Gamma		
Geometric		

Completing Table 0.1 is left as an exercise.

In the chapter on Poisson modeling, we provided heuristic rationale for using the log() function as our link. That is, counts would be non-negative but a linear function inevitably goes negative. By taking the logarithm of our parameter λ we could use a linear predictor and not worry that it can take on negative values. Now we have theoretical justification for this choice, as the log is the canonical link for Poisson data. In the next chapter we encounter yet another type of response, a binary response, which calls for a different link function. Our work here suggests that we will model $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$ using a linear predictor.

[Note that **generalized linear models (GLMs)** differs from **General Linear Models**. The *general* linear model is a statistical linear model with multivariate vectors as responses. For example, each subject in a study may have their height, weight, and shoe size recorded and modeled as a function of age and sex. The response is a vector, Y = (height, weight, shoe size), for each study participant. Age and sex are explanatory variables in the model. The residual is usually assumed to follow a multivariate normal distribution. If the residual is not a multivariate normal distribution, then generalized linear models may be used to relax assumptions about Y and the variance-covariance structure.]

0.41 Exercises

1. For each distribution,

- Write the pdf in one parameter exponential form, if possible.
- Describe an example of a setting where this random variable might be used.

- Identify the canonical link function, and
- Compute $\mu = -\frac{c'(\theta)}{b'(\theta)}$ and $\sigma^2 = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3}$ and compare with known E(Y) and Var(Y).

a) Binary: Y = 1 for a success, 0 for a failure

$$p(y) = p^y(1-p)^{(1-y)}$$

b) Binomial (for fixed n): Y = number of successes in n independent, identical trials

$$p(y) = \left(\begin{array}{c} n \\ y \end{array} \right) p^y(1-p)^{(n-y)}$$

c) Poisson: Y = number of events occurring in a given time (or space) when the average event rate is λ per unit of time (or space)

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

d) Normal (with fixed σ – could set $\sigma = 1$ without loss of generality)

$$f(y;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/(2\sigma^2)}$$

e) Normal (with fixed μ – could set $\mu = 0$ without loss of generality)

$$f(y;\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/(2\sigma^2)}$$

f) Exponential: Y = time spent waiting for the first event in a Poisson process with an average rate of λ events per unit of time

$$f(y) = \lambda e^{-\lambda y}$$

g) Gamma (for fixed r): Y = time spent waiting for the r^{th} event in a Poisson process with an average rate of λ events per unit of time

$$f(y;\lambda) = \frac{\lambda^r}{\Gamma(r)} y^{r-1} e^{-\lambda y}$$

h) Geometric: Y = number of failures before the first success in a Bernoulli process

$$p(y) = (1-p)^y p$$

i) Negative Binomial (for fixed r): Y = number of failures prior to the r^{th} success in a Bernoulli process

$$\begin{aligned} p(y;r) &= \left(\begin{array}{c} y+r-1 \\ r-1 \end{array} \right) (1-p)^y p^r \\ &= \frac{\Gamma(y+r)}{\Gamma(r)y!}(1-p)^y p^r \qquad (0.7) \\ & \qquad (0.8) \end{aligned}$$

j) Pareto (for fixed k):

$$f(y;\theta) = \frac{\theta k^\theta}{y^{(\theta+1)}} \quad \text{for} \quad y \geq k; \theta \geq 1$$

2. Complete Table 0.1 containing your results of the preceding exercises.

0.41.1 Initial Models

Bibliography

John Ashworth Nelder and Robert William Maclagan Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972. doi: 10.2307/2344614. URL http://www.jstor.org/stable/2344614.