Generalized Linear Models

Prof. Sharon McNicholas

STATS 780/CSE 780

Introduction

- In the last "lecture", we took a quick look at linear models.
- Most of what we did was through R.
- I mentioned that what we did was just a "taste", and I recommended Faraway (2014)[a] for further reading.
- In this lecture, we will give generalized linear models (GLMs) a similarly quick look.

[a] Faraway, J.J. (2014). *Linear Models with R*. 2nd edn. Boca Raton: Chapman & Hall/CRC Press.

Introduction contd.

- As with linear models, there are entire courses devoted to GLMs.
- Like last time, most of what we do on GLMs today will be through R.
- This is just a "taste" of GLMs and I recommended Faraway (2016)[a] for further reading.
- I understand that our look at linear models and GLMs will be boring for some of you and too fast for others. . .

[a] Faraway, J.J. (2016). *Extending the Linear Model with R.* 2nd edn. Boca Raton: Chapman & Hall/CRC Press.

Introduction contd.

- I am working up to logistic regression, which will get much more attention.
- Why?
- Two reasons:
  - It is an important topic in data science that you probably have not seen a thorough coverage of.
  - It is our gateway to classification — or at least as I am treating it.
- Note that Assignment 2 focuses on logistic regression.

GLMs

- Again consider $(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \ldots, (\mathbf{X}_n, Y_n)$.

- For a GLM, the response Y is a member of the exponential family and there is a link function defining the relationship between the mean of the response and a linear combination of the predictors.

Exponential Family

- Using the same notation as Faraway (2016), the density of Y from an exponential family has the form

$$p(y \mid \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\},$$

  where

  - θ is the canonical parameter (location),
  - ϕ is the dispersion parameter (scale), and
  - a, b and c are functions.

Exponential Family contd.

- Conveniently,
$$\mathbb{E}[Y] = \frac{\partial}{\partial\theta} b(\theta)$$
and
$$\mathbb{V}\text{ar}[Y] = \frac{\partial^2}{\partial\theta^2} b(\theta) a(\phi).$$

- Let's look at some examples of the exponential family:
  - Gaussian.
  - Poisson.
  - Binomial.

Link Function

- Earlier, we mentioned the idea of having a link function to define the relationship between the mean of the response and a linear combination of the predictors.

- Write $\mathbb{E}[Y] = \mu$.

- To understand the link, write
$$\eta = \beta_0 + \sum_{j=1}^{p} \beta_j X_j, \tag{1}$$
and note that the link is given by
$$\eta = g(\mu).$$

Some Links

- The canonical link has $\eta = g(\mu) = \theta$.
  - For the Gaussian distribution, the canonical link is $\eta = \mu$.
  - For the Poisson distribution, the canonical link is $\eta = \log \mu$ so that $\mu = \exp\{\eta\}$
- Suppose η in (1) is function of $\pi \in [0, 1]$, i.e., we want to relate π to $X_1, \ldots, X_p$. Some common options for the link are:
  - Logit: $\eta = \log[\pi/(1 - \pi)]$.
  - Probit: $\eta = \Phi^{-1}(\pi)$, where Φ is the Gaussian cdf.
  - Complementary log-log: $\eta = \log[-\log(1 - \pi)]$.

Comments

- Now, we will turn to R and look at some data sets.
- Like with linear models, I think that learning in this way is effective in this environment — as part of a quick coverage.
- However, again, I strongly recommend that you do some background reading.
- The first data set we will look at is the famous Challenger data set...

Challenger Data

The US Space Shuttle Challenger exploded during take-off on January 28, 1986. An investigation ensued into the reliability of the shuttle's propulsion system. The explosion was eventually traced to the failure of one of the three field joints on one of the two solid booster rockets. Each of these six field joints includes two O-rings, designed as primary and secondary, which fail when phenomena called erosion and blowby both occur.

The night before the launch, a decision had to be made regarding launch safety. The discussion among engineers and managers leading to this decision included concern that the probability of failure of the O-rings depended on the temperature t at launch, which was forecast to be 31°F. There are strong engineering reasons based on the composition of O-rings to support the judgement that failure probability may rise monotonically as temperature drops.

Challenger Data contd.

No previous liftoff temperature was under 53°F. The task is to predict the number of O-rings that will experience thermal distress for a given flight when the launch temperature is below freezing or, more precisely, when the temperature is 31°F.

The number of O-rings that perished on 23 consecutive launches of the Space Shuttle Challenger were recorded and the data are available as `orings` in the `faraway` package for R.

Based on these data, how many O-rings would you expect to experience thermal distress for a launch temperature of 31°F?