4.6 Generalized Linear Models

Dr. Lauren Perry

Count Data

So far, we've dealt only with qualitative and *continuous* quantitative response variables.

We may also need to work with *discrete* quantitative responses, or *counts*.

The Bikeshare Data

- Response: 'bikers', the number of hourly users of a bike sharing program in Washington DC
- Predictors:
  - `mnth`, month of the year
  - `hr`, hour of the day (0 to 23)
  - `workingday`, indicator for work days (0 if weekend or holiday)
  - `temp`, temperature in Celsius
  - `weathersit`, weather situation: clear; misty or cloudy; light rain/snow; heavy rain/snow

Linear Regression?

Why don't we want use linear regression on count data?

- The linear model
$$Y = X\beta + \epsilon$$
always results in continuous response Y.
  - Since ϵ is continuous, Y must also be continuous.
- Count data is nonnegative, but linear regression outcomes may not be.
- There may also be some data-specific issues that arise.
  - Subsection 4.6.1 has specific examples.

The Poisson Distribution

Suppose a random variables Y takes on nonnegative integer values. If Y follows a Poisson distribution, $Y \sim \text{Poisson}(\lambda)$, then

$$P(Y = k) = \frac{e^{-\lambda}\lambda^k}{k!} \text{ for } k = 0, 1, 2, \ldots$$

- $\lambda > 0$ is the expected values (mean) of Y
  - ... and the variance of Y.
  - That is, $\lambda = \text{E}(Y) = \text{Var}(Y)$

Since it takes on nonnegative integer values, this distribution is typically used to model *counts*.

The Poisson Distribution

Example: Let Y denote the number of users of the bike sharing program (for a set hour of the day, under specific weather conditions, and during a particular month).

If there are 5 users on average per hour under these conditions, we might let $\lambda = 5$ and

$$P(Y = k) = \frac{e^{-5}5^k}{k!}$$

Then the probability of no users in an hour is

$$P(Y = 0) = \frac{e^{-5}5^0}{0!} = e^{-5} \approx 0.007$$

Poisson Regression

We want that mean λ to be able to vary based on our predictor variables: $\lambda(X)$.

That is, we will consider λ as a function of the covariates $X_1 \ldots, X_p$:

$$\log(\lambda(X_1, \ldots, X_p)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

or

$$\lambda(X_1, \ldots, X_p) = \exp\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\right)$$

and $\beta_0, \beta_1, \ldots, \beta_p$ are parameters to be estimated.

- Note: the log of $\lambda(X)$ is linear in X.
  - This ensures that $\lambda(X)$ takes on only nonnegative values (and, by extension, predictions will only take on nonnegative values).

Poisson Regression

To estimate the coefficients $\beta_0, \beta_1, \ldots, \beta_p$, we again use a maximum likelihood approach.

Given n independent observations from the Poisson regression model, the likelihood takes the form

$$l(\beta_0, \beta_1, \ldots, \beta_p) = \Pi_{i=1}^{n} \left(\frac{e^{-\lambda(x_i)} \lambda(x_i)^{y_i}}{y_i!} \right)$$

where $\lambda(x_i) = \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})$

- We estimate the coefficients to maximize this likelihood (to make the observed data as likely as possible).

Poisson Regression on the Bikeshare Data

```r
data(Bikeshare)
contrasts(Bikeshare$hr) = contr.sum(24)
contrasts(Bikeshare$mnth) = contr.sum(12)
mod1 <- glm(bikers ~ workingday + temp + weathersit + hr + mnth,
            data = Bikeshare, family = 'poisson')
```

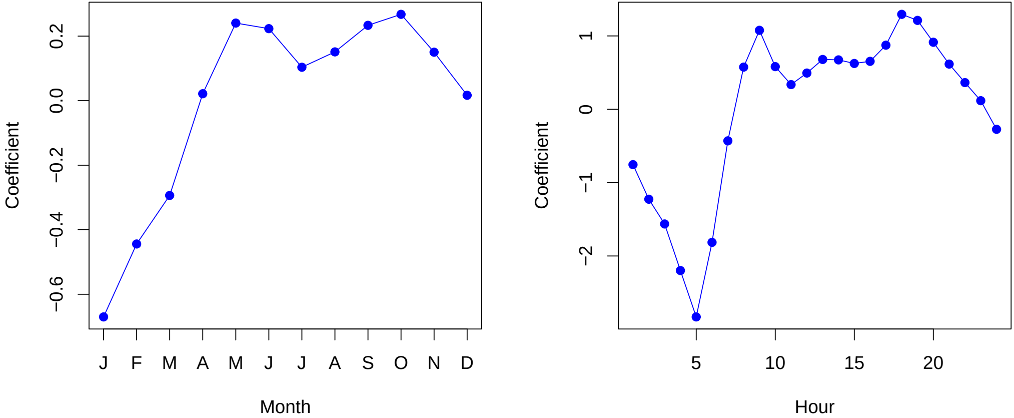
Note: you will see the `contrasts` function in the lab for this chapter. It has to do with the coding of those two variables.

Poisson Regression on the Bikeshare Data

```r
summary(mod1)
```

```
##
## Call:
## glm(formula = bikers ~ workingday + temp + weathersit + hr +
##     mnth, family = "poisson", data = Bikeshare)
##
## Coefficients:
##                            Estimate Std. Error  z value Pr(>|z|)
## (Intercept)                4.118245   0.006021  683.964  < 2e-16 ***
## workingday                 0.014665   0.001955    7.502 6.27e-14 ***
## temp                       0.785292   0.011475   68.434  < 2e-16 ***
## weathersitcloudy/misty    -0.075231   0.002179  -34.528  < 2e-16 ***
## weathersitlight rain/snow -0.575800   0.004058 -141.905  < 2e-16 ***
## weathersitheavy rain/snow -0.926287   0.166782   -5.554 2.79e-08 ***
## hr1                       -0.754386   0.007879  -95.744  < 2e-16 ***
## hr2                       -1.225979   0.009953 -123.173  < 2e-16 ***
```

Poisson Regression on the Bikeshare Data



Poisson vs Linear Regression: Interpretation

A one-unit increase in X_j is associated with a change in $\lambda(X)$ by a factor of $\exp(\beta_j)$.

Example:

- The indicator for cloudy has $\hat{\beta}_{\text{cloudy}} = -0.08$.
- A change in weather from clear (baseline) to cloudy is associated with a change in mean bike usage by a factor of $\exp(-0.08) = 0.923$.
  - That is, on average, 92.3% as many people will use bikes when it is cloudy relative to when it is clear.

Poisson vs Linear Regression: Mean-Variance Relationship

Under the Poisson model, $\lambda = \mathrm{E}(Y) = \mathrm{Var}(Y)$.

- This allows the variance to change with the mean.
  - In the `Bikeshare` data, variability is a lot higher when more people are riding, for example in good weather. This Poisson model accounts for this.
- However, this is also an assumption we make in the Poisson model that may not always hold.

Generalized Linear Models

Discussed three types of regression models:

1. Linear
2. Logistic
3. Poisson

What do these have in common?

Generalized Linear Models

- Each approach uses predictors $X_1, \ldots, X_p$ to predict some response Y.
- Assume that, $Y | X_1, \ldots, X_p$ belongs to a certain family of distributions.
  - Linear regression assumes normal.
  - Logistic regression assumes Bernoulli.
  - Poisson regression assumes Poisson.

Generalized Linear Models

- Each approach models Y as a function of the predictors X.
  - Linear regression

$$\mathrm{E}(Y|X_1,\ldots,X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

  - Logistic regression

$$\mathrm{E}(Y|X_1,\ldots,X_p) = P(Y=1|X_1,\ldots,X_p) = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}{1+\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}$$

  - Poisson regression

$$\mathrm{E}(Y|X_1,\ldots,X_p) = \lambda(X_1,\ldots,X_p) = \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)$$

Link Functions

We can express all of these using a *link function*, η:

$$\eta(\mathrm{E}(Y|X_1, \ldots, X_p)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- For linear regression, $\eta(\mu) = \mu$
- For logistic regression, $\eta(\mu) = \log[\mu/(1 - \mu)]$
- For Poisson regression, $\eta(\mu) = \log(\mu)$

Exponential Family of Distributions

- The normal, Bernoulli, and Poisson distributions are all part of the *exponential family*.
- Other well-known exponential family distributions:
  - Exponential
  - Gamma
  - Negative binomial

Generalized Linear Models

In general, we can perform a regression by modeling Y as coming from any particular member of the exponential family and then transforming the mean of the response so that it is a linear function of the predictors.

Any regression model that follows this approach is a *generalized linear model* (GLM).