Generalized Linear Models

D2K Course Staff

July 25 2019

Introduction

Linear regression models can be pretty useful! However, an ordinary linear regression does not suffice in many cases. Linear regression follows the model:

$$Y = X_i\beta + \epsilon, \epsilon \sim N(0, \sigma^2).$$

This assumes that the response variable is continuous and can, in theory, fall anywhere in the interval $(-\infty, \infty)$. We often ignore this restriction in cases where the linear model works well enough. For example, even though human heights can not be below zero, they exist on a continuous scale and are generally far away enough from their lower bound such that we can ignore this issue and still fit a linear regression model with height as the response variable. Of course, this comes with the understanding that this model should not be used for prediction or inference for cases outside the range of normal human heights (or, more precisely, outside of the range of human heights observed in the data). However, two particular cases where ordinary linear regression does not work is for binary and count data. For one, these data are discrete; thus, using the model $Y = X_i\beta + \epsilon, \epsilon \sim N(0, \sigma^2)$ does not make any sense! Additionally, we may end up with predictions outside the range of possible values, e.g. below 0 for count data. In these cases, we use what is called a generalized linear model. Examples are shown below for the binary and count data casees.

Binary: Logistic Regression

For binary data, using the model $Y_i = X_i\beta + \epsilon, \epsilon \sim N(0, \sigma^2)$ obviously makes no sense. One intiuitive solution might be to use the model:

$$Y_i \sim Bern(p_i)$$
$$p_i = X_i\beta.$$

However, this still runs in to the problem where we can predict probabilities less than 0 or greater than 1. To fix this, we use the logistic regression model:

$$Y_i \sim Bern(p_i)$$
$$\log\left(\frac{p_i}{1-p_i}\right) = X_i\beta.$$

$\frac{p_i}{1-p_i}$ is called the odds ratio. The latter translates mathematically to:

$$p_i = \frac{e^{X_i\beta}}{1+e^{X_i\beta}}.$$

$\log\left(\frac{p_i}{1-p_i}\right)$ falls in the interval $(-\infty, \infty)$ for $0 < p < 1$, solving our issue with predicting probabilities less than 0 or greater than 1.

Below, we fit a logistic regression model predicting whether a melanoma cancer patient has an ulcer based on tumor thickness, taking age and sex in to account. GLMs are generally fit using the `glm()` function in base R. The `family` argument is used to tell the function which type of regression to run - in this case, we want `family - 'binomial'`.

```
library(MASS)
data(Melanoma)
head(Melanoma)
```

```
  time status sex age year thickness ulcer
1   10      3   1  76 1972      6.76     1
2   30      3   1  56 1968      0.65     0
3   35      2   1  41 1977      1.34     0
4   99      3   0  71 1968      2.90     0
5  185      1   1  52 1965     12.08     1
6  204      1   1  28 1971      4.84     1
```

```
model1 <- glm(ulcer ~ age + sex + thickness, family = "binomial", data = Melanoma)
summary(model1)
```

```

Call:
glm(formula = ulcer ~ age + sex + thickness, family = "binomial",
    data = Melanoma)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2198  -0.8287  -0.6813   1.0201   1.8063

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.770495   0.550680  -3.215   0.0013 **
age          0.004499   0.009747   0.462   0.6444
sex          0.331536   0.330579   1.003   0.3159
```

```
thickness    0.430484   0.088727   4.852 1.22e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 281.13  on 204  degrees of freedom
Residual deviance: 234.51  on 201  degrees of freedom
AIC: 242.51

Number of Fisher Scoring iterations: 5
```

Model inference can be done in the same way that it is done in ordinary linear regression; in this model, tumor thickness is a statistically significant prdictor of ulcer presence, conditional age and sex. However, interpreting the specific effects of a variable can be unintuitive. For logistic regression, a coefficient greater than 0 means that an increase in that variable increases p_i, while one less than 0 means that an increase in that variable decreases p_i. The actual quantitative effect of a one unit increase, however, is not linear. What we would typically say is that "a one unit increase in ... increases/decreases the odds $\frac{p_i}{1-p_i}$ of"success" by a factor of e^{β_j}, holding all other variables constant." In this particular example, a one millimeter increase in tumor size increases the odds of having an ulcer by a factor of $e^{0.430484} \approx 1.538$, holding all other variables constant.

Counts: Poisson Regression

In the case of count data, ordinary linear regression can be an appropriate model if the counts are relatively large. However, in the case where we have many 0 counts, we may end up with predicted negative counts. One common option is to use the Poisson regression model in this case:

$$Y_i \sim Pois(\mu_i)$$

$$\log(\mu_i) = X_i\beta.$$

Like above, the logic here is that $\log(\mu_i)$ can be any real-valued number for $\mu > 0$ as required by the Poisson model. The latter formula above can be written as:

$$\mu_i = e^{X_i\beta}.$$

Below, we fit a Poisson regression model predicting the number of snail fatalities given their species, the humditiy and temperature of their enviornment, and the weeks of exposure. This is done using `family = 'poisson'` in `glm()`.

```
data(snails)
head(snails)
```

```
  Species Exposure Rel.Hum Temp Deaths  N
1       A        1    60.0   10      0 20
2       A        1    60.0   15      0 20
3       A        1    60.0   20      0 20
4       A        1    65.8   10      0 20
5       A        1    65.8   15      0 20
6       A        1    65.8   20      0 20
```

```
model2 <- glm(Deaths ~ Temp + Rel.Hum + Exposure + Species, family = "poisson", data = snails)
summary(model2)
```

```

Call:
glm(formula = Deaths ~ Temp + Rel.Hum + Exposure + Species, family = "poisson",
    data = snails)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6461  -0.7860  -0.4230   0.3147   2.0366

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.69349    0.81081   0.855 0.392382
Temp         0.05864    0.01509   3.887 0.000102 ***
Rel.Hum     -0.06777    0.01076  -6.297 3.03e-10 ***
Exposure     1.11708    0.08243  13.553  < 2e-16 ***
SpeciesB     0.85597    0.13182   6.493 8.39e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 448.078  on 95  degrees of freedom
Residual deviance:  64.662  on 91  degrees of freedom
```

```
AIC: 251.45

Number of Fisher Scoring iterations: 5
```

Based on the summary output, all of the predictors in the model are statistically significant. Once again, the interpretation of the coefficients is not straightforward since the effect of each on μ_i is nonlinear. As above, a coefficient greater than 0 means that an increase in that variable increases μ_i, while one less than 0 means that an increase in that variable decreases μ_i. For the specific quantitative effect, we would say that "a one unit increase in ... increases/decreases the mean number of ... by a factor of e^{β_j}, holding all other variables constant." In this particular example, a one week increase in exposure increases the mean number of deaths by a factor of $e^{1.11708} \approx 3.0559$, holding all other variales constant.

Extensions

GLMs can actually be used for regressions under many different types of distributions. The default distributions that can be modeled can be found at `?family`. Additionally, other parametric distributions can be modeled fairly easily using self-written functions - this can generally be done using your preferred optimization algortihm (e.g., `optim` in R, gradient descent) on the likelihood function. (For you math-stat-inclined folks, this is done for exponential family distributions by letting $\theta = X\beta$ for the natural parameter θ of the distribution.)