STATS 111/202

Lecture 8: Components of generalized linear models, examples of link functions

2/3/2025

Ana Maria Kenney

UC Irvine Department of Statistics

Generalized Linear Models

3.1, 3.2, 3.3

Overview

- Finished a review of linear regression under continuous responses and continuous or categorical explanatory variables

- Previously, went over different ways to analyze **tabular data**. Specifically, studying the relationship between 2 categorical variables or 2 categorical variables and a potential confounding variable

- This setting is very limiting. Trying to adjust for several explanatory variables and/or confounders is unreasonable. They do not extend well to larger (more realistic) datasets

Goals with GLMs

- Work with several explanatory variables
- Conduct inference for parameters of interest and control for potential confounders
- Quantify both the direction and strength (magnitude) of association between explanatory variables and the response
- Fit models with quantitative and categorical explanatory variables without having to categorize

Consider a binary response…

- Under the usual regression framework:

$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

    - Where X_{ij} is the ith sample of the jth variable

- Recall, with a binary outcome (success/failure), this follows a Bernoulli distribution:

$$P(Y_i = 1) = p_i$$

- Then we have:

$$E(Y_i) = p_i$$
$$Var(Y_i) = p_i(1 - p_i)$$

Consider a binary response…

- Under the usual regression framework:

$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

  - Where X_{ij} is the ith sample of the jth variable

- Recall, with a binary outcome (success/failure), this follows a Bernoulli distribution:

$$P(Y_i = 1) = p_i$$

- Then we have:

$$\boxed{\begin{gathered} E(Y_i) = p_i \\ Var(Y_i) = p_i(1 - p_i) \end{gathered}}$$

Why is this a problem in our regression framework?

Consider a binary response…

- Under the usual regression framework:

$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

  - Where X_{ij} is the ith sample of the jth variable

- Recall, with a binary outcome (success/failure), this follows a Bernoulli distribution:

$$P(Y_i = 1) = p_i$$

- Then we have:

$$\boxed{\begin{gathered} E(Y_i) = p_i \\ Var(Y_i) = p_i(1 - p_i) \end{gathered}}$$

Why is this a problem in our regression framework?

Non-constant variance!

Consider a binary response…

- Can use the tools for non-constant variance we discussed last week (i.e., weighted least squares)

1. Get an initial estimate for β
2. Obtain $\hat{p}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}$
3. Compute $\widehat{Var}(Y_i) = \hat{p}_i(1 - \hat{p}_i)$
4. Set weights to be $w_i = \frac{1}{\widehat{Var}(Y_i)}$ and refit using these weights
5. Repeat until convergence (when the $\hat{\beta}$ are not changing much from one iteration to another)

Consider a binary response…

- Great! We've dealt with the non-constant variance issue, but there's still something not quite right…

- Our typical regression approach does not restrict the estimated response. What values should $\hat{p}$ follow?

- We do not want negative probabilities, or values greater than 1

- Instead, let's not model $\hat{p}$ directly, but a **function of the probabilities** that can be **inverted** to pull the response back to an acceptable range

$$g\big(E(Y_i)\big) = g(p_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

$$E(Y_i) = g^{-1}\big(g(p_i)\big) = g^{-1}(\beta_0 + \beta_1 X_{i1} + \cdots \beta_p X_{ip})$$

Link function

- Let g() be the natural logarithm

$$\log(p_i) = \beta_0 + \beta_1 X_{i1} + \cdots \beta_p X_{ip}$$

- Then we have

$$p_i = e^{\beta_0 + \beta_1 X_{i1} + \cdots \beta_p X_{ip}} = \exp(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip})$$

- **Did this fix our problem?**
- **No. Now we have nonnegative mean response values, but can still be greater than 1**

GLMs

- Can account for the non-constant variance

- Can model a transformation of the mean of the response and produce fitted values for the response that fit its distribution

GLMs

Three main components:

1. **Random component**
   - Identify an appropriate probability distribution for Y
   - E.g., binomial vs Poisson vs normal
2. **Systematic component**
   - Specify the explanatory variables
   - $X\beta = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$
3. **Link function**
   - Specifies a function g() that relates the population mean μ to the linear combination of predictors
   - $g(\mu_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$
   - Invert this to "go back" to correct scale of responses

GLMs

Interpretation

- Suppose we have binary Y and fit a model

$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

- Here

$$E(Y_i) = P(Y_i = 1)$$

- Then for β_1, it is the change in the probability that the response is equal to 1 for a 1-unit increase in X_1 holding all other covariates constant

GLMs

Interpretation

- Suppose we have binary Y and fit a model

$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

- Here

$$E(Y_i) = P(Y_i = 1)$$

- Then for β_1, it is the change in the probability that the response is equal to 1 for a 1-unit increase in X_1 holding all other covariates constant

Careful!!

- The link function and its inverse will determine how interpretation is done
- For a binary response, will use the link function along with either a RD, RR, or OR to get a useful interpretation of the coefficients

Link functions

Link functions

Identity link function

- Take

$$g(\mu_i) = \mu_i$$

- We have

$$E(Y_i) = \mu_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

- If the random component follows a Normal distribution, then we are back to standard linear regression

Link functions

Logit link function

- Take

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

- We have

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

- Can use to model the log odds for **binary outcomes**

- If μ bounded between 0 and 1 (i.e., proportions), then the log-odds is unbounded $(-\infty, \infty)$

Link functions

Probit link function

- Take

$$g(\mu_i) = \phi^{-1}(\mu_i)$$

Where $\phi(\cdot)$ is the cumulative distribution function of the standard normal

- We have

$$\phi^{-1}(\mu_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

- Resulting in

$$\mu_i = \phi\left(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}\right) = \int_{-\infty}^{X\beta} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

- Also can be used for binary responses

Link functions

Probit link function

- Take

$$g(\mu_i) = \phi^{-1}(\mu_i)$$

Where $\phi(\cdot)$ is the cumulative distribution function of the standard normal

- We have

$$\phi^{-1}(\mu_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

- Resulting in

$$\mu_i = \phi\left(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}\right) = \int_{-\infty}^{X\beta} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

- Also can be used for binary responses

Link functions

Probit link function

- Take

$$g(\mu_i) = \phi^{-1}(\mu_i)$$

Where $\phi(\cdot)$ is the cumulative distribution function of the standard normal

- We have

$$\phi^{-1}(\mu_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

- Resulting in

$$\mu_i = \phi\big(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X$$

- **Also can be used for binary responses**

- Often, logit and probit links produce very similar results
- Logit has slightly fatter tails
- Logit easier to interpret to many due connection with log odds

Link functions

Log link function

- Take

$$g(\mu_i) = \log(\mu_i)$$

- We have

$$\log(\mu_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

- Resulting in

$$\mu_i = e^{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}}$$

- Appropriate when we need $\mu_i > 0$
- Used to model count and rate data using Poisson regression

Link function examples

Example

In R

- Can use the built in glm() function

$$glm(response \sim x_1 + \cdots + x_p, family = familyname\,(link = linkname))$$

- Where

    - Family – which distribution to use to model the response (random component). E.g., gaussian (Normal), binomial (Bernoulli), and poisson

    - Link – function to transform the response (link function). E.g., identity, log, logit, and probit

Example (Midwest house sales, continuous response)

- Using the glm function
- Identity link
- Gaussian family

```
Call:
glm(formula = price ~ sqft, family = gaussian(link = identity),
    data = house)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -81432.946  11551.846  -7.049 5.74e-12 ***
sqft           158.950      4.875  32.605  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 6260433856)

    Null deviance: 9.9109e+12  on 521  degrees of freedom
Residual deviance: 3.2554e+12  on 520  degrees of freedom
AIC: 13260

Number of Fisher Scoring iterations: 2
```

Example (Midwest house sales, continuous response)

- Using the lm function
- Specifies the same model
- Standard linear regression

```
Call:
lm(formula = price ~ sqft, data = house)

Residuals:
    Min      1Q  Median      3Q     Max 
-239405  -39840   -7641   23515  388362 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -81432.946  11551.846  -7.049 5.74e-12 ***
sqft           158.950      4.875  32.605  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 79120 on 520 degrees of freedom
Multiple R-squared:  0.6715,	Adjusted R-squared:  0.6709 
F-statistic:  1063 on 1 and 520 DF,  p-value: < 2.2e-16
```

Example (Titanic survival, binary response)

- Binomial family with identity link

$$\hat{P}(Y_i = 1) = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} = 0.38 - 0.02Age_i$$

- For a 1 unit increase in age, the estimated probability that $Y_i = 1$ decreases by 0.02.

- However, the identity link does not keep predictions restricted to be between 0 and 1

```
Call:
glm(formula = Survived ~ Age, family = binomial(link = "identity"),
    data = titanic_train)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.38376    0.01627  23.583   <2e-16 ***
Age         -0.02464    0.01733  -1.422    0.155
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1186.7  on 890  degrees of freedom
Residual deviance: 1184.6  on 889  degrees of freedom
AIC: 1188.6

Number of Fisher Scoring iterations: 3
```

Example (Titanic survival, binary response)

- Binomial family with logit link

```
Call:
glm(formula = Survived ~ Age, family = binomial(link = "logit"),
    data = titanic_train)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.47528    0.06901  -6.888 5.68e-12 ***
Age         -0.10868    0.07462  -1.456    0.145
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1186.7  on 890  degrees of freedom
Residual deviance: 1184.5  on 889  degrees of freedom
AIC: 1188.5

Number of Fisher Scoring iterations: 4
```

Announcements

- Extra practice from textbook: 2.23, 2.27 part a and c, 2.33, 2.37, and 2.39
- HW #3 will be due next Sunday, February 9^{th}
- Midterm review sheet will be posted this Wednesday