Generalized Linear Models

Sean Hellingman ©

Regression for Applied Data Science (ADSC2020)

shellingman@tru.ca

Winter 2025

THOMPSON RIVERS UNIVERSITY

Topics

2 Introduction
3 Why GLMs
4 Generalization
5 Exponential Family
6 Variance
7 Link Function
8 R Basics
9 Additional Information
10 Exercises and References

Introduction

- Linear models are extremely useful but require very strong assumptions to be valid.

- Linear models can be *generalized* to work for a wider variety of tasks.

- Each model still has required assumptions but they are generally easier to satisfy.
  - More flexibility in the models.

Linear Regression Model Assumptions I

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_j X_j + \epsilon$$

1. **Linearity**
   - The relationship between the dependent and independent variable(s) needs to be linear.
2. **Normality** (multivariate normal for multiple independent variables)
   - In linear regression, all variables must be normally distributed.
3. **Homoscedasticity** (constant variance)
   - The variation about the regression line is constant for all values of the independent variable(s).
4. **Independence**
   - There is little or no multicollinearity in the data (independent variables are too highly correlated with each other).

Linear Regression Model Assumptions II

- Based on the assumptions the error term (ϵ):

$$\epsilon \sim N(0, \sigma^2)$$

- We can use the residuals e_i from our estimated linear regression models to check these assumptions (related to the error term).
- Do we need anything else?

Why Generalized Linear Models

- In many cases the variance depends on the explanatory variables
  - Variance can naturally depend on the mean.
  - Sometimes the relationship can be very complicated.
- The additive relationship assumed by linear models can be unrealistic.
- The response variable no longer follows a normal distribution.

Example 1

- Load the *dental.csv* data into R.
- Take some time to get to know that data.
- Estimate the following linear regression model:

  $$\text{DMFT} = 1 + \text{Sugar} + \text{Indus}$$

- Will this model pass the diagnostics?

Generalization I

- Linear Regression (matrix notation):
  - $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\sigma^2 I})$
  - $E(\boldsymbol{Y}|\boldsymbol{X}) = \boldsymbol{\mu} = \boldsymbol{X}\beta$
- Generalized Linear Regression (matrix notation):
  - $\boldsymbol{Y} \sim$ Exponential Family
  - $E(\boldsymbol{Y}|\boldsymbol{X}) = \boldsymbol{\mu} = g^{-1}(\boldsymbol{X}\beta)$

Generalization II

- $\boldsymbol{Y} \sim$ Exponential Family

- The **Exponential Family** is a *family* of distributions that contains many widely used distributions:
    - Normal
    - Binomial
    - Poisson
    - Gamma
    - ...

- $E(\boldsymbol{Y}|\boldsymbol{X}) = \boldsymbol{\mu} = g^{-1}(\boldsymbol{X}\beta)$

  Link Function

- $g\left[E(\boldsymbol{Y}|\boldsymbol{X})\right] = g[\boldsymbol{\mu}] = \boldsymbol{X}\beta$

Exponential Family I

- It is now assumed that the response follows a distribution from the *natural exponential family.*
  - Not the same as the exponential distribution.
- Density:

$$f_\theta(y) = exp[\{y\theta - b(\theta)\}/a(\phi) + c(y, \phi)] \tag{1}$$

  - ϕ: dispersion parameter
  - θ: canonical parameter (function of β)
  - a,b,c: functions

Normal Distribution

$$
\begin{aligned}
f_\mu(y) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] \\
&= \exp\left[\frac{-y^2+2y\mu-\mu^2}{2\sigma^2} - \log(\sigma\sqrt{2\pi})\right] \\
&= \exp\left[\frac{y\mu-\mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \log(\sigma\sqrt{2\pi})\right]
\end{aligned}
$$

$$\theta = \mu,\ b(\theta) = \theta^2/2 \equiv \mu^2/2,\ a(\phi) = \phi = \sigma^2$$

$$c(\phi, y) = -y^2/(2\phi) - \log(\sqrt{\phi 2\pi}) \equiv -y^2/(2\sigma^2) - \log(\sigma\sqrt{2\pi})$$

Poisson Distribution

$$
\begin{aligned}
f(y;\lambda) &= \frac{\lambda^y e^{-\lambda}}{y!} \\
&= \exp\left\{\frac{y\log\lambda - \lambda}{1} - \log y!\right\}
\end{aligned}
$$

$$\theta = \log\lambda,\ b(\theta) = e^{\theta} \quad a(\phi) = 1$$

Exponential Family II

Distribution	θ	$a(\theta)$	ϕ
Binomial(n, π)	$\ln(\frac{\pi}{1-\pi})$	$n\ln(1+e^{\theta})$	1
Poisson(μ)	$\ln(\mu)$	e^{θ}	1
Normal(μ, σ^2)	μ	$\frac{1}{2}\theta^2$	σ^2
Gamma(μ, ν)	$-\frac{1}{\mu}$	$-\ln(-\theta)$	$\frac{1}{\nu}$
Inverse Gaussian(μ, σ^2)	$-\frac{1}{2\mu^2}$	$-\sqrt{-2\theta}$	σ^2
Negative Binomial(μ, κ)	$\ln(\frac{\kappa\mu}{1+\kappa\mu})$	$\frac{1}{\kappa}\ln(1-\kappa e^{\theta})$	1

Variance of the Exponential Family

- The variance of Y is a function of the mean (see next slide):

$$Var(Y) = a(\phi)V(\mu)$$

- And the mean is a function of the explanatory variables.

- Therefore, the variance is also a function of the explanatory variables $\rightarrow$ **heteroskedasticity**.

Exponential Family Variance Functions

Distribution	E(y)	$V(\mu) = \frac{\text{Var}(y)}{\phi}$
Binomial(n, π)	$n\pi$	$n\pi(1-\pi)$
Poisson(μ)	μ	μ
Normal(μ, σ^2)	μ	1
Gamma(μ, ν)	μ	μ^2
Inverse Gaussian(μ, σ^2)	μ	μ^3
Negative Binomial(μ, κ)	μ	$\mu(1+\kappa\mu)$

Example 2

- Simulate observations various distributions from the previous slides and create histograms of your observations.
- Can you think of any examples where any of these distributions could be better applied than the normal distribution?
- How does changing the parameters change the shapes of the distributions?

Link Function

- The *additive effect* of the explanatory variables on the response is assumed on some transformation of the mean.
- The **link function** is used to perform this transformation.
  - Can be selected for the specific task (logistic regression).
  - There are recommended combinations of link functions and distributions.

$$E(\boldsymbol{Y}|\boldsymbol{X}) = \boldsymbol{\mu} = g^{-1}(\boldsymbol{X}\beta) = g^{-1}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...)$$

$$g\left[E(\boldsymbol{Y}|\boldsymbol{X})\right] = g[\boldsymbol{\mu}] = \boldsymbol{X}\boldsymbol{\beta} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...$$

Common Link Functions

Link Function	g(μ)	Common link for
Identity	μ	Normal
Log	$\ln(\mu)$	Poisson
Power	μ^p	Gamma ($p = -1$) Inverse Gaussian ($p = -2$)
Square Root	$\sqrt{\mu}$	
Logit	$\ln(\frac{\mu}{1-\mu})$	Binomial

Maximum Likelihood Estimation

- **Maximum Likelihood Estimation (MLE)** is a method of estimating parameters from an assumed probability distribution.
  - **This is the primary estimation method used to estimate GLMs.**

- *What parameter(s) values are most likely to have generated the observations.*

- Estimates are obtained by maximising a likelihood function based on an assumed distribution.

- It is assumed that the distribution that the data are drawn from is known.

`glm()`

- To estimate a GLM in R:
  - `GLModel <- glm(response ~ Var.1 + Var.2 + ...  + Var.j, family = distribution.name(link = "default.link"), data = data)`

- You can use the `summary()` and `predict()` functions as you would for linear models.

- *We will speak about interpreting the individual results for task specific models as go through them.*

Example 3

- Use the *dental.csv* data and the `glm()` function to estimate the linear model from Example 1.
- Are there other distributions you think will improve the model?

Other Useful Distributions I

- Lognormal Distribution:
  - Stock prices & real estate prices
- Gamma Distribution:
  - Insurance risk
- Weibull Distribution:
  - Time to failure
- Beta Distribution:
  - Fire risk

Other Useful Distributions II

- Geometric Distribution:
  - Number of trials until the first success
- Negative Binomial Distribution:
  - Count response variable (over-dispersed)
- Logistic Distribution:
  - Binary response variable
- Poisson Distribution:
  - Count response variable

Exercise 1

- Take some time to read about the possible distribution choices of GLMs.
- Try to apply some of these models to examples where we have struggled to satisfy the assumptions of linear regression.

References & Resources

1. De Jong, P., & Heller, G. Z. (2008). *Generalized linear models for insurance data*. Cambridge University Press.

- `glm()`
- `family()`