Continuous Generalized Linear Models

Sean Hellingman ©

Regression for Applied Data Science (ADSC2020)

shellingman@tru.ca

Winter 2025

THOMPSON RIVERS UNIVERSITY

Topics

Introduction

- Generalized linear models (GLMs) can be used in more situations than linear regression
  - More flexibility in the response variable.
  - Relaxation of some of the assumptions required.
- As with linear regression, we can assume a numeric response variable.

Numeric Response Variables

- Assuming a continuous response variable that is no longer normally distributed.
- *Methods are useful for positively skewed data.*
- The *gamma distribution* and the *inverse Gaussian distribution* are members of the exponential family and offer more flexibility than the normal distribution.

Gamma Distribution

- The **gamma distribution** is a flexible continuous distribution with two parameters used in many areas.

- One of the PDFs of the gamma distribution:

$$f(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta} \tag{1}$$

    - $k > 0$ shape parameter
    - $\theta > 0$ scale parameter

Inverse Gaussian Distribution

- The **inverse Gaussian distribution** (Wald distribution) is a flexible continuous distribution with two parameters used in many areas.
- The PDF of the inverse Gaussian distribution is:

$$f(x) = \sqrt{\frac{\lambda}{2\pi x^3}} exp \left[-\frac{\lambda(x-\mu)^2}{2\mu^2 x} \right] \tag{2}$$

  - $x > 0$
  - $\mu > 0$
  - $\lambda > 0$

Gamma Regression Model

- When constructing gamma regression, the mean μ, is modeled in terms of explanatory variables:

$$g(\mu) = \boldsymbol{X}\beta$$

- Using the inverse link function:

$$\mu^{-1} = \boldsymbol{X}\beta$$

- *The log link function is also commonly used.*

Gamma Regression in R

- Estimating a Gamma regression model in R is very similar to a linear regression model:
  - `GammaModel <- glm(response ~ Var.1 + Var.2 + ... + Var.j, family = gamma(link="inverse"), data = data)`

- Estimated using MLE.
- `link = "inverse"` is included by default.

Example 1

- Import the *Cars93* dataset from the MASS package
- Examine the histogram of the Max.Price variable.
- Estimate the following gamma regression models:

$$(Max.Price)^{-1} = MPG.highway + Horsepower + DriveTrain$$
$$ln(Max.Price) = MPG.highway + Horsepower + DriveTrain$$

- Interpret the results.

Interpreting Coefficients from Gamma Regression

- Using the `summary()` function, **the coefficients express how a one unit change in the explanatory variable changes the inverse of the expected value of the response** (inverse link)
- Using the `summary()` function, **the coefficients express how a one unit change in the explanatory variable changes the log of the expected value of the response** (log link)
- *Generally, it is easier to interpret the coefficients generated with the log link function.*

Inverse Gaussian Regression Model

- When constructing inverse Gaussian regression, the mean μ, is modeled in terms of explanatory variables:

$$g(\mu) = \boldsymbol{X}\beta$$

- Using the recommended link function:

$$\mu^{-2} = \boldsymbol{X}\beta$$

- *The log link function is also commonly used.*

Inverse Gaussian Regression in R

- Estimating a Gamma regression model in R is very similar to a linear regression model:
  - `IGModel <- glm(response ~ Var.1 + Var.2 + ... + Var.j, family = inverse.gaussian(link = "1/mu^2"), data = data)`

- Estimated using MLE.
- `link = "1/mu^2"` is included by default.

Example 2

- Import the *Cars93* dataset from the MASS package
- Estimate the following inverse Gaussian regression models:
  $$(Max.Price)^{-2} = MPG.highway + Horsepower + DriveTrain$$
  $$ln(Max.Price) = MPG.highway + Horsepower + DriveTrain$$
- Interpret the results.

Interpreting Coefficients from Inverse Gaussian

- Using the `summary()` function, **the coefficients express how a one unit change in the explanatory variable changes the of the expected value of *1 over the response variable squared*** (1/mu^2 link)

- Using the `summary()` function, **the coefficients express how a one unit change in the explanatory variable changes the log of the expected value of the response** (log link)

- *Generally, it is easier to interpret the coefficients generated with the log link function.*

Assumptions

- We have relaxed the normality of the response and homoscedasticity assumptions.
- **The following assumptions still apply for gamma/inverse Gaussian regression:**
  1. The response variable is bounded by zero.
  2. There is a linear relationship between the continuous predictor variables and the transformed dependent variable (through the link function).
  3. There is **no** multicollinearity of the explanatory variables.

Checking Assumptions

- *To check linearity, follow the visualization steps outlined in Binary Logistic Regression and Models for Count Data.*
  - If violated:
    - Try to transform explanatory variables to create a linear relationship (polynomials).
    - May be able to use regression splines.
- The no multicolinearity assumption can be checked using the `vif()` function from the *car* package.
- Cook's distance may be used to examine models for potential outliers (values over 0.5 and 1.0).

Likelihood Ratio Test

- ***Likelihood ratio tests*** *can be used to assess the goodness of fit of two competing statistical models.*

- As both models are estimated with MLE, they can be compared.

- In R:
  - `library("lmtest")`
  - `lrtest(Model1,Model2)`

- The null hypothesis is that Model1 is *as good as or better* than Model2.
  - *Model2 is usually a more complex model*

Information Criteria

- We can use AIC and BIC to compare **gamma and inverse Gaussian** models.
- Remember, lower values of AIC and BIC are considered to be better.
- BIC has a higher penalization for the number of parameters than AIC.
- In R:
  - `AIC()`
  - `BIC()`

Example 3

- Ignoring the assumptions for now, which of the four estimated models is the best (inferential objective).

Predictions

- All of the models may be used to make predictions.
  - `predict(GLMModel, type = "response")`

- We may use *caret* or the *boot* package for cross-validation.

Example 4

- Use the *boot* package to compare the average prediction accuracy of the four models we have estimated so far.

Comments on Continuous Generalized Linear Models

- Dealing with negative response variables may be challenging.
  - May be able to truncate or transform the response variable.
  - May be able to use more complex or generalized distributions.
  - May be able to apply non-parametric regression techniques.

- Generalized linear models should be useful for many situations you encounter.

Exercise 1

- Run the appropriate diagnostics for the *best* model identified in Example 3.
- Does this model violate any of the assumption, and if so what can be done to try and correct the problem?

Exercise 2

- Take some time to estimate some regression models with continuous response variables that do not appear to follow the normal distribution, test the assumptions, make predictions, and test prediction accuracy.

References & Resources

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

- `glm()`
- `family()`
- `AIC()`
- `cv.glm()`