

Gamma Regression

Consider a continuous dependent variable that is positive-valued, such as a length of a hospital stay, time waiting or the cost of a bill. This type of data is continuous in nature and oftentimes skewed and a normal approximation does not hold.

The type of data above presents a constant **Coefficient of Variation (CV)**, that is:

$$\sqrt{\text{var}(Y_i)} \text{ over } E(Y_i) = \sigma$$

The identity induces a quadratic variance function:

$$\text{var}(Y_i) = \sigma^2 E(Y_i)^2$$

To model such data a number of approaches proposed in the literature have proved useful, including **Log-Normal models** and **Gamma Regression models**

1. Log Normal models - the log transformation followed by a classical linear model is fairly successful in modeling this type of data. This approximation works best when the scale parameter (σ) is small. Indeed, the log transformed model has approximately constant variance:

$$\log(Y) \approx \log(\mu) + (Y - \mu) \left\{ \frac{\Delta \log(y)}{\Delta y} \right\}(\mu) = \log(\mu) + \left\{ \frac{Y - \mu}{\mu} \right\}$$

$$\text{Var}(\log(Y)) \approx \left\{ \frac{\text{Var}(Y)}{\mu^2} \right\} = \left\{ \frac{\sigma^2 \mu^2}{\mu^2} \right\} = \sigma^2$$

2. Gamma Regression - the Gamma regression keeps the outcome on the original scale. If one wants to work on the original scale the framework of the generalized linear model proves very fruitful.

Gamma Distributions

The **Gamma** family is a very flexible family of distributions with support on the positive axis. The family is indexed by two parameters. One way to parameterize which is used in SAS and focused on in this lecture is called the **mean parameterization**.

A variable is said to follow the Gamma distribution if the density has the form:

$$f_Y(y) = \frac{1}{\Gamma(v)y} \left\{ \frac{yv}{\mu} \right\}^v \exp\left(-\frac{yv}{\mu}\right), 0 < y < \infty$$

where

$$\Gamma(v) = \int_0^\infty x^{v-1} \exp(-x) dx$$

The mean and variance of a Gamma distributed variable are:

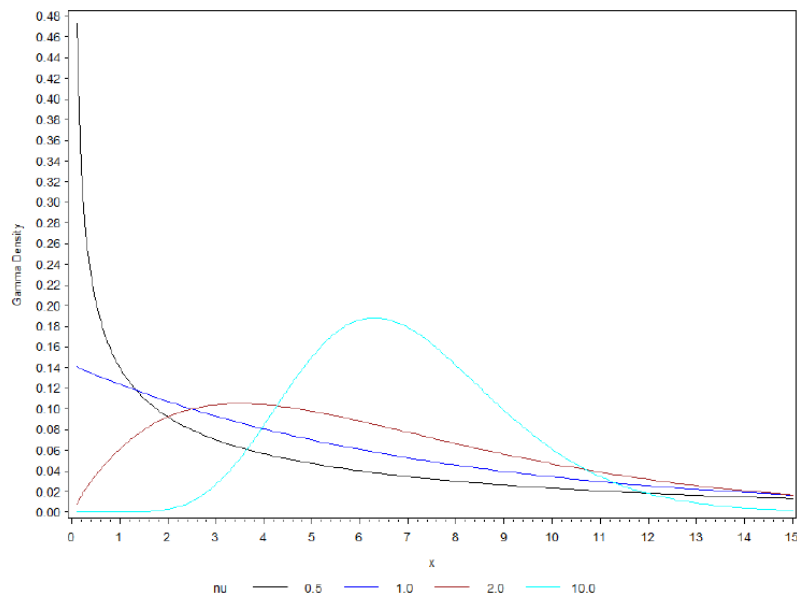
$$E(Y) = \mu$$

$$\text{var}(Y) = \frac{\mu^2}{v} = \frac{\sigma^2}{\mu^2}, \sigma^2 = \frac{1}{v}$$

The parameter v is called the **scale parameter**, and (v / μ) is called the **rate parameter**. The inverse of the scale parameter $(1/\sigma^2)$ is called the **dispersion parameter**.

Properties of the Gamma Distribution

1. The shape of the density is controlled by v
 1. Regardless of the value of v the densities are skewed to the right
2. When $v < 1$ the Gamma distribution is exponentially shaped and asymptotic to both vertical and horizontal axes
3. A Gamma distribution with scale parameter $v = 1$ and mean parameter μ is the same as an exponential distribution with mean parameter μ . In SAS we can actually test whether $v = 1$
4. When $v > 1$ the Gamma distribution assumes a unimodal but skewed shape. The skewedness reduces as the value of v increases. As v tends to ∞ the Gamma distribution begins to resemble a normal distribution.
2. The chi-square distribution is a special case of the Gamma. A chi-square distribution with n degrees of freedom is the same as a Gamma with $(v = \{n \over 2\})$ and $(\mu = n)$
3. The Gamma is a flexible life distribution model that may offer a good fit to some sets of failure data. However, it is not widely used as a life distribution model for common failure mechanisms.
4. The Gamma does arise naturally as the time-to-first failure distribution for a parallel system with components having lifetimes distributed exponentially. If there are n components in the system and all components have exponential lifetimes with mean $(1/\gamma)$, then the total lifetime has Gamma distribution $v = n$ and $(\mu = n/\gamma)$
5. The Gamma distribution is widely used in engineering, science and business.



The above is an example of different shapes of the Gamma distribution with different values of ν and $\mu = 7$

Gamma as a Generalized Linear Model

As with all Generalized Linear Models, to specify the Gamma regression as a GLM we need to specify the link and variance function besides the distribution of the response.

1. Variance function is $V(\mu) = a \cdot \mu^2$
2. The most commonly used link function are the Log ($g(\mu) = \log(\mu)$) and inverse ($g(\mu) = 1 / \mu$). The default in SAS is inverse
 - With inverse link, a change in the coefficients will induce an opposite change in the expected value of the response, with log link the opposite is true
 - The inverse link does not map the expected value μ into the whole real line; therefore we have to be careful when we interpret parameters

Interpretation

Assume we have a dichotomous exposure X ($1 = \text{yes}$, $0 = \text{no}$). If the log link is used we can describe the regression as:

$$\log(\mu(X)) = \alpha + \beta \cdot X$$

The coefficient β is interpreted in terms of **Mean Ratio (MR)**

$$MR = \frac{\exp(\alpha + \beta)}{\exp(\alpha)}$$

The increase in logarithm of means for one unit increase in X .

With an inverse link a change in the coefficients will induce an opposite change in the expected value of the response.

$$E\{1 \text{ over } \mu(X)\} = \alpha + \beta X$$

The coefficient β is interpreted in terms of **Inverse Mean Difference (IMD)**

$$IMD = \{1 \text{ over } \mu_{X=1}\} - \{1 \text{ over } \mu_{X=0}\} = \alpha + \beta - \alpha = \beta$$

One unit increase in X causes an increase of β in the inverse of means.

With log link the interpretation of the parameters is similar to the logit models; however the odds are replaced by the means, and the OR is replaced with mean ratios:

$$\log\left(\frac{\mu_i}{\mu_{i'}}\right) = \beta(X_i - X_{i'})$$

The expected response is multiplied by $\exp(\beta)$ for each unit change in X

With inverse power metric a change in the coefficients will induce an opposite change:

$$\frac{1}{\mu_i} - \frac{1}{\mu_{i'}} = \beta(X_i - X_{i'})$$

The inverse of the expected response changes by β for each unit change in X . If $\beta > 0$ then the mean decreases with an increase in X , while if $\beta < 0$ then the mean increases with an increase in X .

Measuring Goodness of Fit

The log-likelihood for a given model M :

$$l(\mu, \nu, y) = \sum_{i=1}^n \left\{ \nu \left[-\frac{y_i}{\mu_i} - \ln(\mu_i) \right] - \ln(\Gamma(\nu)) + \nu \ln(\nu y_i) - \ln(y_i) \right\}$$

The log-likelihood for a saturated model:

$$l(y, \nu, y) = \sum_{i=1}^n \left\{ \nu [-1 - \ln(y_i)] - \ln(\Gamma(\nu)) + \nu \ln(\nu y_i) - \ln(y_i) \right\}$$

Thus the deviance is:

$$D(y, \hat{\mu}) = 2(l(y, \nu, y) - l(\mu, \nu, y)) = -2\nu \sum_{i=1}^n \left[\ln\left(\frac{y_i}{\mu_i}\right) - \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right]$$

In Gamma regression the scaled deviance and the scaled Pearson chi-square are measures of goodness of fit statistics. For the Gamma distribution $(\sigma^2 = \frac{1}{\nu})$. If ν is a known priori for a model M that has p parameters and n observations, the scaled deviance and scaled Pearson chi-square are:

$$\frac{D}{\sigma^2} \text{ and } \frac{X^2}{\sigma^2} \approx \chi^2_{n-p}$$

That is reject the model if:

$$\frac{D}{\sigma^2} > \chi^2_{n-p, 1-\alpha}$$

Where σ is estimated from the data. Thus to compare two nested models, with p and 1 parameters:

$$\frac{\{D_{M_2} - D_{M_1}\}}{\sigma^2} \approx \chi^2_{p-1}$$

These estimates can be very unstable; however, one reasonable approach is to estimate it in the largest model we are willing to consider then use that estimate in model selection.

Estimating the Dispersion Parameter

For a Gamma distributed variable Z , with Mean = $E(Z) = \mu$ and Variance = $\text{Var}(Z) = \frac{1}{\mu^2}$, with a response Y_i for subject i we can estimate variance as:

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^n (Y_i - 1)^2 = \frac{1}{n-p} \sum_{i=1}^n \left(\frac{Z_i}{\mu_i} - 1 \right)^2 = \frac{1}{n-p} \sum_{i=1}^n \left(\frac{Z_i - \mu_i}{\mu_i} \right)^2$$

For a normal linear regression a similar result holds. The scale parameter in PROC GENMOD is the standard deviation.

Exploratory Data Analysis

To check the assumptions of a Gamma regression and to help choose an appropriate function form and link function you can plot the predictors X vs $\log(E(Y))$ and $1/E(Y)$. The more appropriate link would appear linear.

With ordinal predictors you would compute the raw means of the response at each level of the predictor while for continuous outcomes you can aggregate values of predictors and treatment as ordinal.

SAS Code

```
/******  
/** Gamma Model - Log normal          **/  
/******  
title1 'Gamma regression - Saturated model';  
title3 'Inverse Link';  
proc genmod data=claims;  
class policyholderage cargroup ageofcar ;  
model lcost=policyholderage|cargroup|ageofcar @2/dist=normal ;  
weight number;  
run;  
  
/******  
/** Gamma Model - Log Link            **/  
/******
```

```
/******  
title1 ' Gamma regression - Saturated model';  
title3 'Log Link';  
proc genmod data=claims;  
class policyholderage cargroup ageofcar ;  
model cost=policyholderage|cargroup|ageofcar /dist=gamma link=log ;  
weight number;  
run;
```

Revision #26

Created 11 April 2023 18:16:08 by Elkip

Updated 13 April 2023 13:54:54 by Elkip