Reading: GLM.Basics
Model: Source text
Problem Type: Predict auto claim severity using a GLM

Given

y	Target variable = loss cost
x_1	Driver age (predictor)
x_2	Marital status (predictor)
log	Link function
Gamma	Distribution

<= Model specification for GLM software, input along with a data set of observations.

<= We assume the loss cost after accounting for the predictors is random and follows a Gamma distribution.

Coefficient	Parameter
5.8	β_0 (Intercept)
0.1	β_1 (Coefficient for driver age)
-0.15	β_2 (Coefficient for marital status)
0.3	ϕ (Dispersion parameter)

<= GLM Software output

Find

a.) Predict the average claim severity for:
- i.) A 25-year old married driver
- ii.) A 35-year old unmarried driver

b.) Calculate the variance of the loss cost for:
- i.) A 25-year old married driver
- ii.) A 35-year old unmarried driver

Solution

To begin we need to understand the types of predictor variables used in the GLM. To do this, look at the model output.

Marital status is clearly a categorical variable as there isn't a continuous range of marital statuses. Looking at the model output, since there is only one coefficient (β_2) for marital status, we infer marital status is a binary variable, so either 1 or 0.

We're dependent on the question to specify which marital status corresponds to 0 and 1 respectively. Since it isn't explicitly called out, assume since most people are unmarried, that 0 = unmarried and 1=married. (This also matches with the logic of 1 = True and 0 = False.)

Next, driver age could be treated as either a continuous or discrete/categorical variable as we typically measure age in a whole number of years. Since the GLM output only has one coefficient for driver age (β_1) we infer age is a continuous variable as otherwise there would be a coefficient $\beta_{1,i}$ for each age in the data set.

Now we understand the GLM output, we can set up the GLM equation as follows: $g(\mu_i) = \ln(\mu_i) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2$

Here we're using the natural logarithm for the log-link function g.

Now it's a matter of plugging in the numbers and then inverting the link function

a.) i.) $g(\mu_i)$ = 5.8 + 0.10 * 25 + -0.15 * 1 **<= Remember this driver is married so marital status = 1**

= 8.15

Inverting the link function by exponentiating gives

μ_i = 3,463.38 **<= This is the predicted average loss cost for a claim for the set of married 25-year old drivers**

a.) ii.) $g(\mu_i)$ = 5.8 + 0.10 * 35 + -0.15 * 0

= 9.3

Inverting the link function by exponentiating gives

μ_i = 10,938.02 **<= This is the predicted average loss cost for a claim for the set of unmarried 35-year old drivers**

Notice how we could also write this as $\mu_i = e^{\beta_0} \cdot e^{\beta_1 \cdot x_1} \cdot e^{\beta_2 \cdot x_2}$

In a.)i.) above this becomes μ_i = 330.30 * 12.182 * 0.861

We can split this apart as:

330.30 is the "base rate" – the average severity for the whole book of business/data set

12.182 is the factor for a driver aged 25

0.861 is the factor for a married driver

We can further interpret the results of a.) as follows:

a.) i.) The severity distribution for the set of married 25-year old drivers follows a Gamma distribution with μ = 3,463.38 and ϕ = 0.3

a.) ii.) The severity distribution for the set of unmarried 35-year old drivers follows a Gamma distribution with μ = 10,938.02 and ϕ = 0.3

Notice in both cases we have ϕ = 0.3. This is because ϕ is assumed to be constant across the entire data set.

b.) We now have fully specified Gamma distributions for part a.) so we can calculate the variance as ϕ * V(μ), which for a Gamma distribution is $\phi * \mu^2$

b. i.)	Variance =	0.3 * 3,463.38 ^2	=	3,598,498.37
b. ii.)	Variance =	0.3 * 10,938.02 ^2	=	35,892,079.26

The higher-risk driver (determined by the average claim severity, μ_i) has a higher variance than the lower risk driver despite ϕ being constant.