Final Course Project

Linear models & Generalized linear models

Arnau García & Maria Lee

An assessment presented for the course of
Linear Models & Generalized Linear Models 2023/24 (MESIO)

Facultat de Matemàtiques i Estadística
Universitat Politècnica de Catalunya
Date of submission: 18/Jan/2024
Professor: Víctor Peña Pizarro

Contents

1 Introduction

No one can doubt the importance of forests and trees for human beings. Forests have been on our planet forever, long before us. And although in recent times they have suffered due to human activities and over-exploitation, it is undeniable that forests are still part of everyone's daily life.

In this work we have tried to solve the following question: can we predict and model some characteristics of a forest given other properties of the forest? In order to answer this question the first step was to search for a good database of trees. For this purpose we contacted scientists from CREAF [1] (*Centre de Recerca Ecològica i d'Aplicacions Forestals*). Víctor Granda, a data scientist at CREAF, advised us in this regard, and thanks to him we found the incredible database of the *Laboratori Forestal Català* [2].This is a very complete database, maintained by Víctor himself, on the forests of Catalonia. The amount of observations and variables available is very large, and we have decided to work only with the forests of the Vallès Oriental and Occidental regions (see these forests in Figure 1).

In order to use several of the concepts learned during the course, we have decided to model two variables: a continuous variable and a categorical variable. The continuous variable will be the diameter at the breast height (DBH), and our objective will be to analyze whether a generalized linear model following a gamma distribution is a better choice for modeling than the classical linear normal model. The categorical variable will be the dominant genus per basal area, in this case we want to use a multinomial model to predict the predominant genus in a forest taking into account its characteristics.

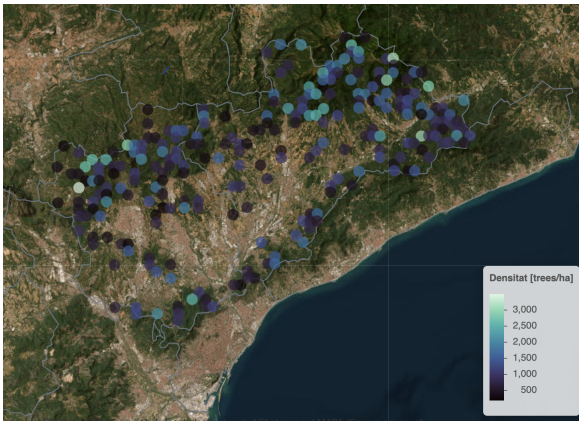

Figure 1: Map with the forests of our dataset.

[1] See the website on `https://www.creaf.cat/es`

[2] See the database on `https://laboratoriforestal.creaf.cat/nfi_app/`

2 Descriptive analysis

Our final dataset has a total of 254 observations, for which we selected 26 variables out of the many that the complete dataset contains. However, these variables are not totally independent from each other and, in fact, some of them seem to be heavily correlated. For this reason, we ran some analyses and finally decided to keep just the ones that made more sense for each model, ensuring that all the predictors were independent from one another. This "pre-selection" will be implied in all of the following sections, where the starting variables in the models are already selected so that they are uncorrelated to each other.

In Table 1 we include a description of all the variables that were finally included in some of the models, and in Table 2 we include a frequency table of all the categorical variables.

Variable name	Description
`dbh`	Diameter at breast high (DBH). It is the diameter of the trunk of a standing tree. Units: cm.
`basal_area`	Basal area is the cross-sectional area of trees at breast height. Units: m^2/ha.
`basal_area_genus_dominant`	Dominant genus per basal area. The categories are `Quercus`, `Pinus` and `Other`.
`density_dec_dominant`	Dominant density of Sclerophyll, Conifer or Deciduous. The categories have the same name.
`basal_area_dec_dominant`	Dominant basal area of Sclerophyll, Conifer or Deciduous. The categories have the same name.
`density`	Density of standing trees. Units: trees/ha.
`coords_latitude`	Latitude in the WGS84 projection.
`admin_region`	Administrative region. In our case they are Vallès Oriental and Vallès Occidental.
`clim_prec_year`	Yearly precipitation (2023). Units: L.

Table 1: Description of the variables in our dataset.

`basal_area_genus_dominant`	`density_dec_dominant`	`admin_region`
Pinus : 117	Deciduous : 30	Vallès Occidental : 100
Quercus : 121	Conifer : 53	Vallès Oriental : 154
Others : 16	Sclerophyll : 171	

Table 2: Frequencies of the categorical variables of our dataset.

3 Methods

3.1 Gamma regression model

One of our main interests in this work has been to see if there are models other than the classical normal linear models that work well for modeling continuous outcomes. Although we have seen in some works (see [3]) the use of GLM with the gamma distribution, its use is not very common and this type of models are not used as frequently as the normal linear models or other GLM. In this section we will be using a GLM with the Gamma distribution to model the diameter at the breast height of the forests. First we will develop an exploratory analysis of our data. Then, we will build a normal linear model and a GLM using the gamma distribution and we will compare both. As it is always important to carry out the work rigorously, we have included an appendix with the demonstration that we can use the gamma distribution in a GLM (see Appendix A).

3.1.1 Exploratory analysis

In the process of selecting variables for our model we have consulted different papers in which mortality or DBH growth models of trees are carried out (see [1], [4], or [5]). We have used these papers as a benchmark to know which variables should be included in the model. After this literature review, we have concluded that the `basal_area` and `density` variables are key. Thus, what we have done is check how the `dbh` behaves with respect this variables. The reader can observe in the scatterplots exposed in the Figure 2 that the relationship of the variable `dbh` with the variables `density` and `basal_area` is not nearly linear. In addition, a sort of curvature can be observed in the data, and also how the variance changes as the x_i. This suggest that a heteroskedastic model (a model where the variance depends on x_i) may be suitable. The reader can also see in this scatterplots the straight line which is the one corresponding to the simple linear regression of the respective variables.

We have also exposed in Figure 2 two boxplots relating the `dbh` with the `basal_area_genus_dominant` and with the `basal_area_dec_dominant` variables, respectively. Although there are no clear differences between the DBH of these categorical variables, there are some distinctions (such as between pinus and quercus) that may be useful in our model.

Of course, we have also developed this exploratory analysis with the rest of the variables in our dataset, but the have decided to expose these results as they are the most interesting and most related to the model we will see. In addition, in the exploration of the other variables we have been able to see that they did not offer us any type of information on the DBH, therefore the other variables have not been included in the model.

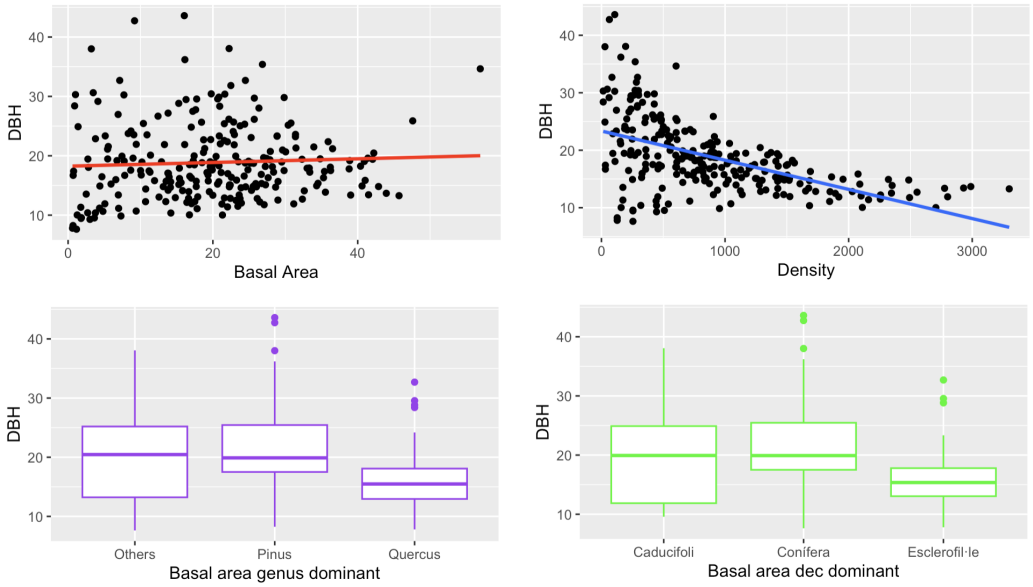

Figure 2: Scatter plots and boxplots of the `dbh` with respect other variables of interest.

3.1.2 Fitting and interpreting the model

Once developed the exploratory analysis and discarded the variables that do not contribute any information to our problem and that would penalize our model, we play with the following variables: `basal_area`, `density`, `basal_area_genus_dominant`, `basal_area_dec_dominant`, `topo_altitude_asl`, `admin_region`, `admin_municipality`, `clim_prec_year`. For our purposes, and due what we have read in the literature, it is mandatory to include `basal_area` and `density` variables.

As we can see in the Figure 2, this is a hard problem. And it seems that the normal linear model is not the best option. Observe that if we use a GLM with the Gamma distribution and using the exponential as a response function we have

$$E(y_i|x_i) = h(\eta_i) = \exp\left(x_i^t\beta\right) = \frac{\alpha}{\nu}, \quad Var(y_i|x_i) = \frac{\alpha}{\nu^2}. \tag{1}$$

Note that the reader can see this notation well introduced in the Appendix A. This is, we have a heteroskedastic model. Thus, it may be a good idea to try to model the DBH using a GLM with the Gamma distribution and the exponential as a response function (i.e. the logarithm as a link function). What we are going to do, is to fit the best GLM possible with our variables. As we commented, we want to add, at least, the `density` and `basal_area` variables to our model. Thus we start with these variables as a benchmark and we add more variables by means of a best subsets algorithm for model selection (feasible due the final set of variables contains eight variables), using as a criterion the Akaike's information criterion (AIC). By doing this we have obtained the model the summary of which can be found in the Table 3.

The AIC of the model obtained is 1366.54, while the AIC of the normal linear model fitted using

Variable	Estimate	Std. Error	t value	p-value
Intercept	2.931	$5.629 \cdot 10^{-2}$	52.069	$< 2 \cdot 10^{-16}$
`density`	$-4.860 \cdot 10^{-4}$	$2.882 \cdot 10^{-5}$	-16.864	$< 2 \cdot 10^{-16}$
`basal_area`	$2.182 \cdot 10^{-2}$	$1.657 \cdot 10^{-3}$	13.172	$< 2 \cdot 10^{-16}$
`basal_area_genus_dominantPinus`	$2.027 \cdot 10^{-1}$	$8.995 \cdot 10^{-2}$	2.254	0.0251
`basal_area_genus_dominantQuercus`	$1.158 \cdot 10^{-1}$	$8.351 \cdot 10^{-2}$	1.387	0.1666
`basal_area_dec_dominantConífera`	$-2.021 \cdot 10^{-1}$	$8.527 \cdot 10^{-2}$	-2.370	0.0185
`basal_area_dec_dominantEsclerofil·le`	$-1.936 \cdot 10^{-1}$	$7.342 \cdot 10^{-2}$	-2.637	0.0089

Table 3: Summary of the GLM obtained.

the same variables is 1448.619. Then, our generalized linear model is much better in terms of AIC. Moreover, we have fitted the best normal linear model possible using exactly the same criterion and best subsets algorithm than the used for fitting the GLM, and we have obtained a model with the same variables except `basal_area_genus_dominant`. Notwithstanding, for the models with this variables (all the above except `basal_area_genus_dominant`) the AIC for the normal linear model is 1446.229, while the AIC of the GLM is 1367.24, which is still considerably lower.

Now, it is time to interpret the model obtained. In the Table 3 we have the different p-values for each estimated coefficient. We can see that all the variables are significant at a 0.05 level except the `basal_area_genus_dominant` for the Quercus category, but the Pinus category is significant at 0.05 level, so the addition of this variable is admissible. According to the result obtained the dispersion parameter of the model is $\hat{\phi} = 1/\hat{\alpha} = 0.0397$. In order to interpret the coefficients obtained we have to take into account that, since we are using the logarithm as a link function, we are dealing with a log-linear model. Using this we have that:

- The `density` coefficient is negative, meaning that the bigger the density of trees in a forest the lower the DBH. In addition, if we increase in 1000 trees per hectare the density of a given forest, the DBH decreases by a factor of $\exp(-4.860 \cdot 10^{-4} \cdot 1000) = 0.615$.
- The `basal_area` is positive, meaning that the bigger the basal area the bigger the DBH. In addition, if we increase in 10 square meters per hectare the basal area, the expected DBH increases $\exp(0.02182 \cdot 100) = 1.24$ times (it is, an increase of a 24%).
- The `basal_area_genus_dominant` coefficients for Pinus and for Quercus are positive, meaning that Pinus and Quercus has larger DBH than the baseline category (which is Other types of trees).
- The `basal_area_dec_dominant` coefficients for *Conífera* and *Esclerofile* are negative, meaning that these type of trees has smaller DBH than the baseline cateogry, which is *Caducifoli.*

Finally, we can predict the DBH values and develop the plots shown in the Figure 3. As the reader can see, these plots explain much better how the data behaves compared with the straight lines exposed in the Figure 2. Moreover, of course, we have carried out a residuals analysis. In the case of the GLM we have obtained symmetric deviance residuals, which makes us think that it is a good choice. With respect to the other residuals, although the results are not perfect, and there are

some outliers, everything indicates that they are (the linear normal and the GLM) sufficiently valid models.

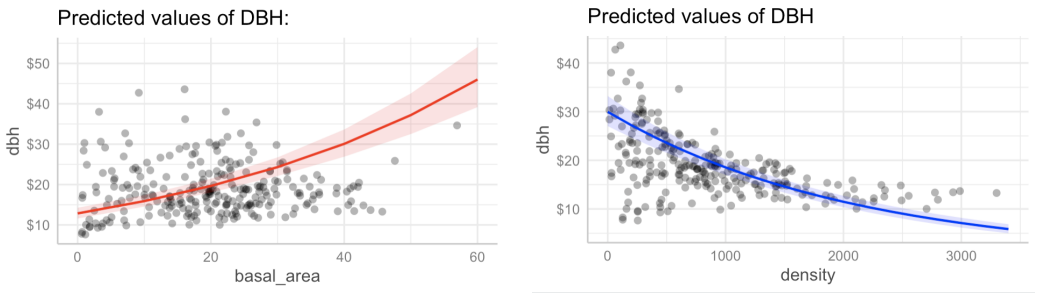

Figure 3: Predicted values of the DBH by the generalized linear model fitted.

3.2 Multinomial logistic regression

The aim of this section is to predict the trees' dominant genus per basal area, where the three possible outcomes are `Pinus`, `Quercus` and `Others`. In order to do so, we split randomly our original dataset with a proportion 20/80 into two different datasets, using the latter to train our model.

3.2.1 Exploratory analysis

First, we will carry out an exploratory data analysis, to see if our model will be good at classifying the outcomes. We have 10 initial variables: `basal_area_genus_dominant`, `admin_region`, `density_dec_dominant`, `coords_longitude`, `coords_latitude`, `admin_municipality`, `dbh`, `density`, `basal_area`, and `clim_prec_year`. We know that some of the important predictors may be `basal_area`, `dbh`, `density` and `admin_region`, so we will check those first, as can be seen in Figure 4. The first impression we get is that it is not going to be very easy to classify these genus, as it seems they share quite a lot of characteristics. The variable `admin_region`, that tells us whether the tree is in the Vallès Oriental or Vallès Occidental, seems that might help us a bit more.

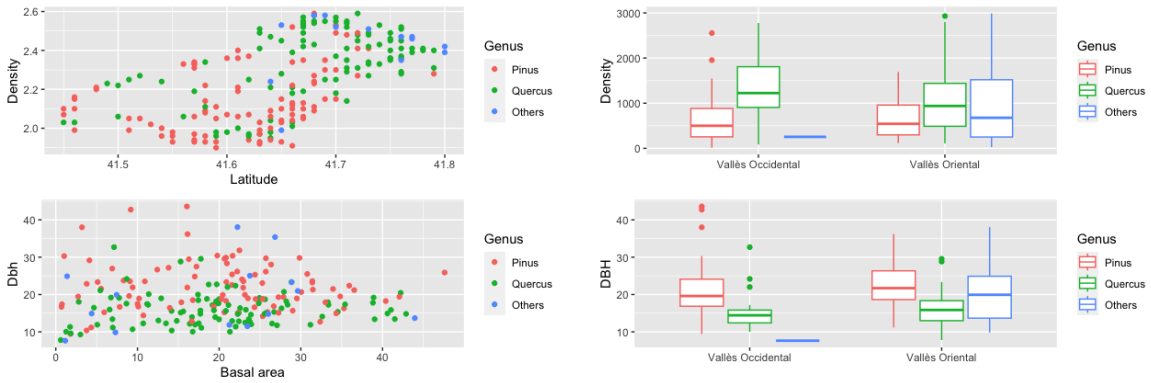

Figure 4: Scatterplots and boxplots represented by dominant genus of several variables.

3.2.2 Fitting and interpreting the model

Following, our goal is to find a model fit given the variables we chose from the dataset. Since we have already used the best subsets for the Gamma regression in the previous section, to do it differently, we now implement the usual step algorithm, with AIC as the criterion for choosing the best model fit. We use the backward selection and we get that our model will include the following variables:

$$\texttt{basal\_area\_genus\_dominant} \sim \texttt{basal\_area} + \texttt{density\_dec\_dominant} + \texttt{density} + \texttt{coords\_latitude} + \texttt{admin\_region} + \texttt{clim\_prec\_year}$$

where the coefficients for the model are given in Table 4.

	(Intercept)	Basal area	DEC:Conifer	DEC:Sclerophyll
`Quercus`	132.395	-0.163	-15.304	1.274
`Others`	-611.634	-0.067	-6.315	-48.619
	Density	Latitude	Vallès Oriental	Precipitation
`Quercus`	0.003	-3.428	1.236	0.013
`Others`	-0.002	13.798	-0.188	0.055

Table 4: Coefficients for the multinomial model to predict the dominant genus per basal area

Keeping in mind that our baseline category for the dominant genus is `Pinus`, we can interpret them as follows:

- The coefficient for `basal_area` is negative for both `Quercus` and `Others`. This means that higher basal area values are associated with a decreased probability of the tree being `Quercus` or `Others` relative to `Pinus`, with less probability for `Quercus` relative to `Pinus` than `Others` relative to `Pinus`.
- The coefficient for `coords_latitude` is negative for `Quercus` and positive for `Others`. This means that higher latitude values are associated with a decreased probability of the tree being `Quercus` relative to `Pinus` and with an increased probability of the tree being `Others`.
- The coefficient for `admin_region` corresponding to Vallès Oriental is positive for `Quercus` and negative for `Others`. This means that if a tree is in Vallès Oriental, it has a higher probability of it being `Quercus` relative to `Pinus`, and lower probability of it being `Others`.
- The coefficient for `density_dec_dominant:Conifer` is negative for both `Quercus` or `Others`, which means that for conifer dominant forests there is lower probability of the trees being `Quercus` and `Others` relative to `Pinus`. For `density_dec_dominant:Sclerophyll`, however, it is positive for `Quercus` and negative for `Others`.

We can see the representation for `density` in Figure 5, according to `admin_region`, where we can see that for higher values of density, there is an increased probability of the tree being `Quercus`, and a decreased probability of being `Others`, which confirms what we see in Table 4. For `Pinus`

and `Others` it seems that the probabilities are higher for those in Vallès Occidental and viceversa for `Quercus`.

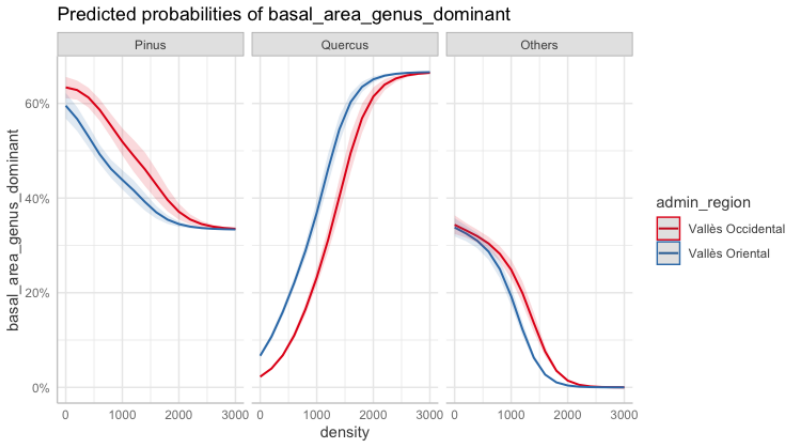

Figure 5: Predicted probabilities for dominant genus by density and administrative region.

Lastly, we are going to use this model to try and predict the dominant genus of our test dataset. To see how well our model performs, we compute its confusion matrix and we compare its accuracy to the no information rate (NIR).

Actual / Predicted	`Pinus`	`Quercus`	`Others`
`Pinus`	19	3	0
`Quercus`	3	20	0
`Others`	2	1	2

Since the category for `Others` has much less observations than the other two, it makes sense that it is the one that performs worse. However, considering that the characteristics for the genus overlap considerably, we conclude that our model performs quite well for `Quercus` and `Pinus`. We get an accuracy of 82%, much higher than the value for the NIR, which is 48%. The difference between our model and the NIR one is significant, since the p-value obtained testing their difference is very small and we can reject the null hypothesis that they both perform equally.

4 Conclusions and future research

In this work, we addressed the question of whether it is possible to predict and model certain characteristics of forests based on other properties. We focused on modeling the DBH as a continuous variable using a GLM with the Gamma distribution. The exploration revealed a non-linear relationship between DBH and variables like tree density and basal area. The GLM with the Gamma distribution outperformed the traditional linear model in terms of AIC, indicating its suitability for capturing the complexities of the data. The model identified key predictors such as tree density, basal area, and dominant genus per basal area, shedding light on the factors influencing DBH in the studied forests. On the other hand, it should be noted that modeling this problem using a normal linear model and transformations could work well. In fact, we have tried using various transformations with logarithms and have obtained better results than those of the normal linear model without transformations. Thus, a comparison between models with the addition of transformations would be interesting for future research.

Additionally, we employed multinomial logistic regression to predict the dominant genus per basal area, classifying trees into categories such as `Pinus`, `Quercus`, and `Others`. The model incorporated variables like basal area, density, latitude, region, and precipitation. The results highlighted the significance of certain predictors in determining the dominant genus. Notably, higher basal area favored `Pinus`, while latitude and region played crucial roles in distinguishing between `Quercus` and `Others`. The model demonstrated promising predictive capabilities, especially for `Quercus` and `Pinus`, achieving an accuracy of 82%, well above the no information rate.

In conclusion, our analyses provide valuable insights into the relationships between forest characteristics and variables such as DBH and dominant genus. The chosen modeling approaches offer effective tools for understanding and predicting these features, contributing to the broader understanding of forest ecosystems.

Finally, we wanted to comment that the final work we have carried out has differences with the work proposal we made. In the work proposal we proposed to model the diameter growth of the trees (as is done in works such as [2], [4] or [1]), the problem is that in these models the temporal factor came into play. We thought that this type of study, with data that evolve over time, was not the right one for this work. Nevertheless, we think that it is an interesting topic and that it remains for us to address in the future.

References

[1] Patricia Adame, Jari Hynynen, Isabel Canellas, and Miren del Río. Individual-tree diameter growth model for rebollo oak (quercus pyrenaica willd.) coppices. *Forest Ecology and Management*, 255(3-4):1011–1022, 2008.

[2] Danaza Mabvurira and Jari Miina. Individual-tree growth and mortality models for eucalyptus grandis (hill) maiden plantations in zimbabwe. *Forest Ecology and Management*, 161(1-3):231–245, 2002.

[3] Victoria KY Ng and Robert A Cribbie. Using the gamma generalized linear model for modeling continuous, skewed and heteroscedastic outcomes in psychology. *Current Psychology*, 36(2):225–235, 2017.

[4] Marc Palahí, Timo Pukkala, Jari Miina, and Gregorio Montero. Individual-tree growth and mortality models for scots pine (pinus sylvestris l.) in north-east spain. *Annals of Forest Science*, 60(1):1–10, 2003.

[5] Nirmal Subedi and Mahadev Sharma. Individual-tree diameter growth models for black spruce and jack pine plantations in northern ontario. *Forest Ecology and Management*, 261(11):2140–2148, 2011.

A Gamma distribution in a GLM

Now, our goal is to demonstrate that we can use the Gamma distribution in the GLM theory. This is, assuming that the distribution of y_i given x_i is a Gamma distribution with parameters α, ν, we want to proof that:

1. The Gamma distribution is within the exponential family of distributions.
2. There exists an invertible, twice-differentiable function h such that $E(y_i|x_i) = h(x_i^t\beta)$.

Let's start with the first point. We will be working with the *exponential dispersion model*, this model states that if the distribution of y_i given x_i is within the exponential family of distributions, then the probability density (or mass) function can be written as

$$f(y_i) = \exp\left(\frac{\eta_i y_i - A(\eta_i)}{\phi} + c(y_i, \phi)\right). \tag{2}$$

We are going to proof that we can write the density function of $Gamma(\alpha, \nu)$ in the shape of (2). Notice that the probability density function of a $Gamma(\alpha, \nu)$ is:

$$f(y_i) = \frac{y_i^{\alpha-1}\exp(-\nu y_i)\nu^\alpha}{\Gamma(\alpha)} \quad \text{for } y_i > 0, \quad \alpha, \nu > 0. \tag{3}$$

Now, we rewrite the previous expression:

$$\begin{aligned} f(y_i) &= \exp\left(-\nu y_i + (\alpha-1)\log y_i + \alpha\log\nu - \log\Gamma(\alpha)\right) \\ &= \exp\left(\frac{\frac{-\nu}{\alpha}y_i - (-\log\nu)}{\frac{1}{\alpha}} + (\alpha-1)\log y_i - \log\Gamma(\alpha)\right) \\ &= \exp\left(\left(\frac{\eta_i y_i - A(\eta_i)}{\phi} + c(y_i,\phi)\right)\right). \end{aligned}$$

Where $\eta_i = -\nu/\alpha$, $\phi = 1/\alpha$, $A(\eta_i) = -\log(-\eta)$ and $c(y_i, \phi) = \frac{-\log\phi}{1/\phi} + (1/\phi - 1)\log y_i - \log\Gamma(1/\phi)$. And then, we have obtained that, indeed, the Gamma distribution is within the exponential family of distributions.

Now, we are interested in demonstrate the second part, but this is easy using the relationship between the GLM conditions and the *canonical response function*. We have that

$$E(y_i|x_i) = h(\eta_i) = \frac{\partial A(\eta_i)}{\partial \eta_i} = -\frac{1}{\eta_i}.$$

And then the second point is also satisfied. In conclusion, we can use the Gamma distribution for develop a generalized linear model.

Notice that is not necessary to use as a response function the canonical. There are other response functions, for instance $h = \exp()$, that can be used. Although the canonical response function has good theoretical properties and is easy to find using the connection between the GLM conditions, in practice is interesting to check if other response functions can lead to better models. Due to the nature of the data we are working with, the exponential response function is more appropriate.