# **Drills with R on generalized linear models**

Suraj Eswaran

University Of The Cumberlands

2024 Spring – Statistics For Data Science (MSDS -531 – A01) – First Bi-Term

Professor Dr. Danny T. Barnes

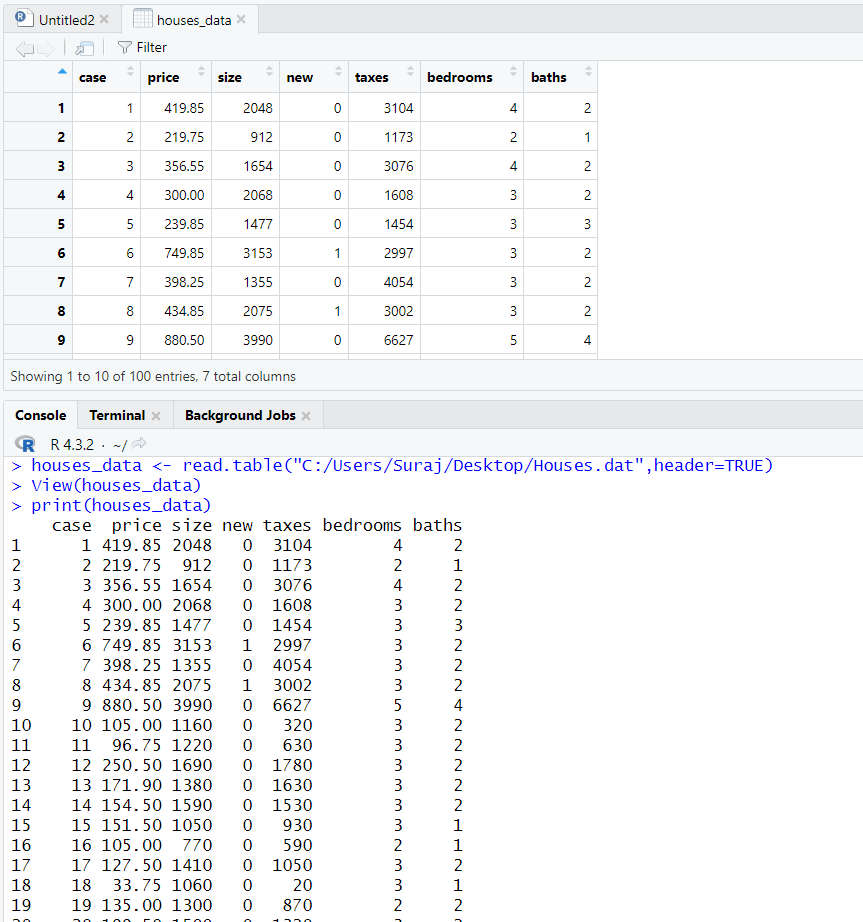
02/11/2024

**For the Houses data at Index of Datasets consider Y = selling price, x1 = tax bill (in dollars), and x2 = whether the house is new:**

In order to read the dat file, we have used read.table function where we need to read houses.dat file which consists of case, selling price, size, new, taxes bills, number of bedrooms and bathroom, setting header to be equal to “TRUE” (Thulin, 2021, p. 120).



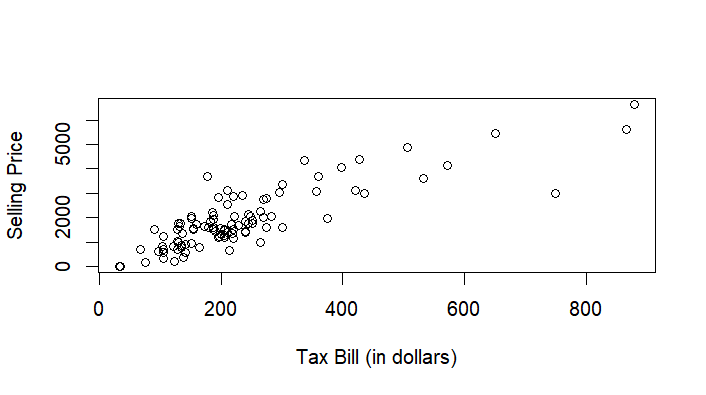
Viewing the data can be done with the help of View and print function, where we will be able to see all the variables and their values(Thulin, 2021, p. 38).



**Form the scatter plot of y and x1. Then answer, does the normal GLM structure of constant variability in y seem appropriate? If not, how does it seem to be violated?**

We have plotted scatter plot with the help of plot function, setting x1 as tax bill and y to be selling price from the dataset. The dots in the scatter plots shows us that whether there is a correlation that is existing between two points. Here, we have taken tax and selling price for plotting scatter plot, where we could see a slight strong, positive, relationship. As the selling price increases, the tax also increases.



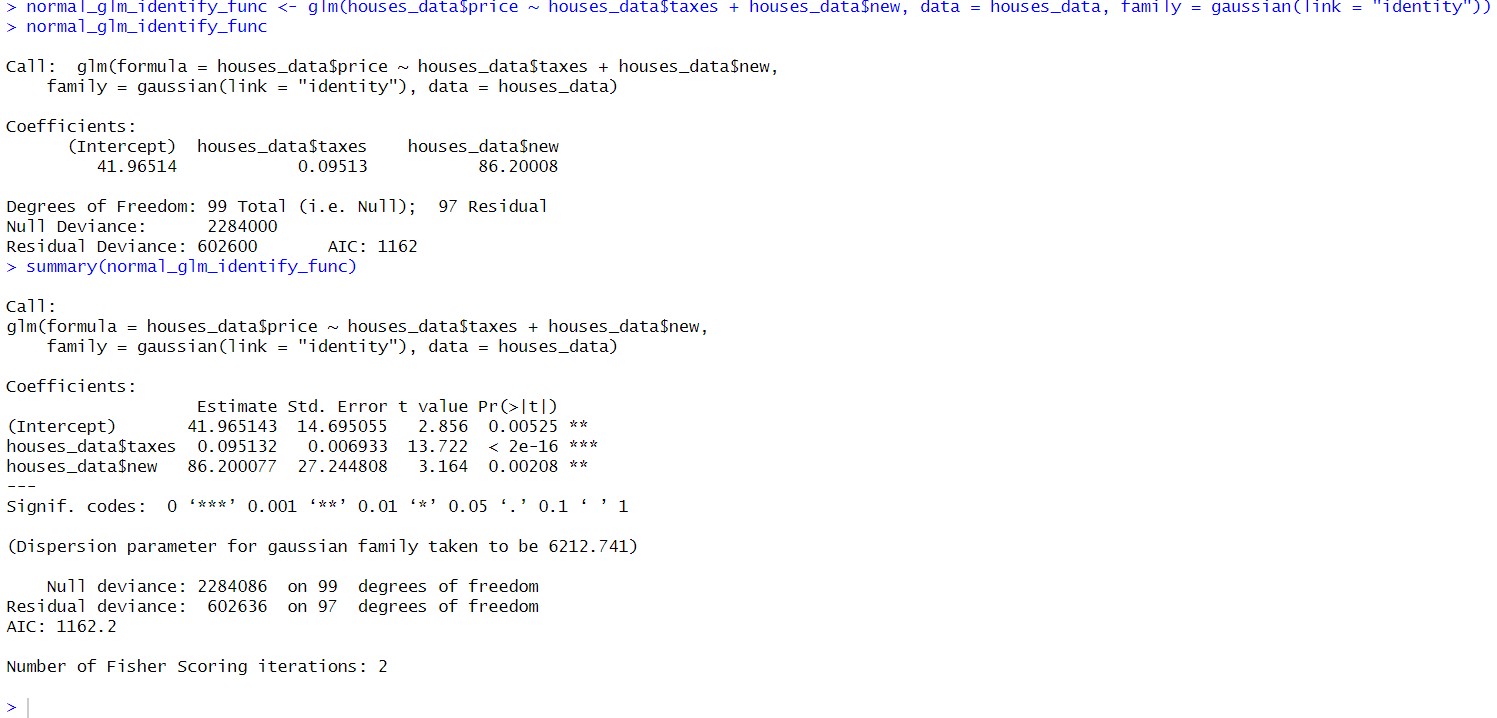


In order to identify whether normal GLM structure of constant variability in Selling Price(y) seem appropriate, we could refer whether the spread of Selling Price(y) across Tax Bill(x1) is constant or not from the scatter plot. As we could see that spread of Selling Price appears to increase as tax bill increases, thus it might not be adequately captured by a normal Generalized Linear Model (GLM) structure as there will be the assumption of constant variance may be violated. In such cases, a standard Generalized Linear Model (GLM) may not be appropriate, as GLMs assume constant variance. From the scatter plot, we could infer that Gamma-based GLM will be a better model as compared normal GLM.

**Using the identity link function, fit the**

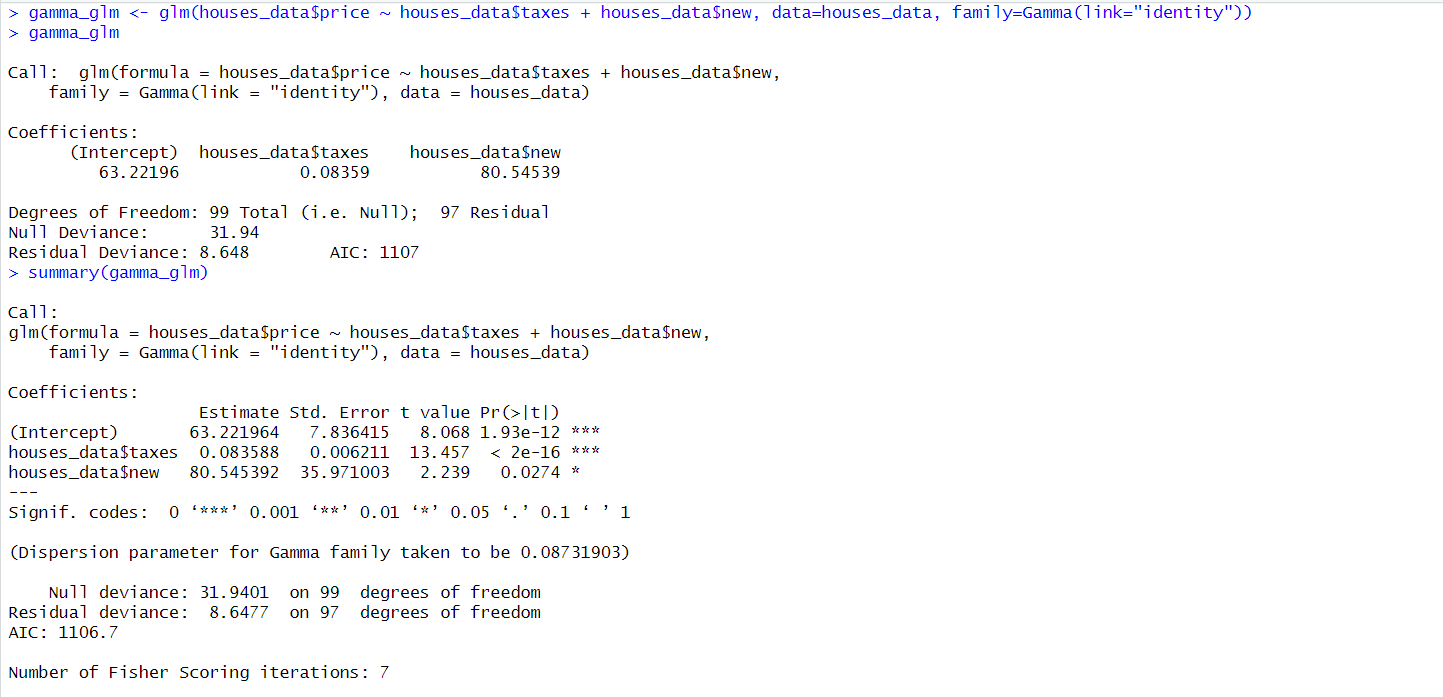
**normal GLM**

In order to fit a normal GLM in R, we have used glm() function where y(Selling Price) is response variable and x(taxes bill + new) and family=gaussian indicates that we are using GLM method with Gaussian (normal) distribution. Once the model is fitted successfully, we have utilized summary() function where we could see result of coefficients, standard errors, p-values, and t-values.

****

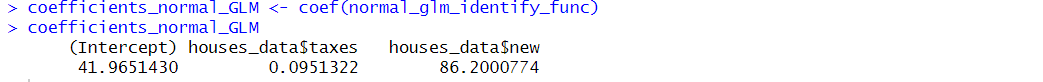
**gamma GLM**

In order to fit a gamma GLM in R, we have used glm() function where y(Selling Price) is response variable and x(taxes bill + new) and family=gaussian indicates that we are using GLM method with Gamma (link as identity) distribution. Once the model is fitted successfully, we have utilized summary() function where we could see result of coefficients, standard errors, p-values, and t-values.

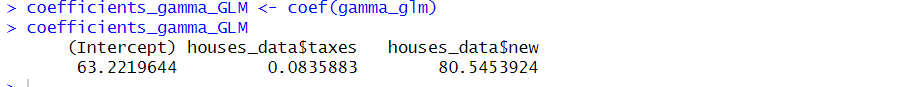
****

**For each model, interpret the effect of x2.**

For normal GLM, the coefficients of x2(new) shows the value of difference between mean selling price of new house with mean selling price of old price from our normal model. If the difference is positive, then the selling price of new house is greater than old price, whereas if its difference is negative, then the selling price of old house is greater than new price. We have used coef() function in order to extract the coefficients of a normal(Gaussian) GLM model. The coefficients variable will have intercept between two predictor variables with the coefficients for each predictor variable used in the model. From the result from coef() function, by seeing the difference value, you can interpret that the mean selling price for new house is higher than that of an mean selling price for old house.



For gamma GLM, the coefficients of x2(new) shows the value of different between mean selling price of new house with mean selling price of old price from our gamma model. If the coefficient value is positive, then the selling price of new house is greater than old price, whereas if the coefficient value is negative, then the selling price of old house is greater than new price. In order to extract the value of coefficient value from gamma GLM model, we have used coef() function, thus giving us the coefficients associated with each predictor variable in our model, including the intercept.

****

**For each model, describe how the estimated variability in selling prices varies as the mean selling price varies from 100 thousand to 500 thousand dollars.**

Coefficient of variation is calculated as the ratio of standard deviation with mean (Shechtman, 1970). In normal GLM, if the variability remains constant, it means the Coefficient of variation remains constant. Therefore, as the mean selling price increases , the standard deviation of the selling prices stays the same, there won’t be any impact on variability. In other words, the variability relative to the mean does not change over this range(100,000 – 500,000).

In a Gamma GLM, the variance of the distribution is proportional to the square of the mean. In a Gamma GLM, k(shape parameter) and p(scale parameter) are estimated from the data. Typically, p is fixed while k is estimated. However, the key point is that the variance increases as a power function of the mean. This indicates that if there is an increase in mean value of selling price, then the coefficient of variation will increase, thus there will be an increase in variability.

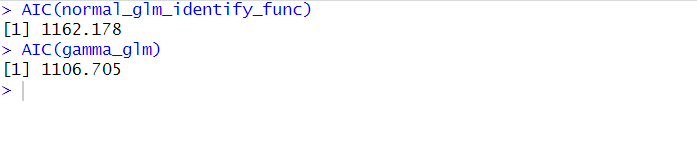
It's important to keep in mind these differences in assumptions and implications when choosing between normal and gamma GLMs, particularly when considering how the variability in the response variable relates to the predictor variables.

**Which model is preferred according to AIC?**

To determine which model is preferred for this data, we are comparing with the help of Akaike Information Criterion (AIC), where it is recommended to select the model with the lowest AIC value. We have calculated AIC value for both normal GLM model and Gamma GLM with the help of AIC method.

For Normal GLM model, we have received a value of1162.178 and for Gamma GLM model we have value of 1106.705. Akaike Information Criterion is a method for validating whether the model is fit or not. IF we have lower AIC scores, we could understand that model has a better balance between model fit and complexity.

Since Gamma GLM model has the lower AIC value, it is the preferred model according to AIC as compared to Normal GLM model. Lower AIC values indicate better balance between goodness of fit and model complexity (Narisetty, 2020). Therefore, Gamma GLM model is the preferred model.



## References

Thulin, M. (2021). Modern Statistics with R: From wrangling and exploring data to inference and predictive modelling. BoD-Books on Demand.

Shechtman, O. (2013). The coefficient of variation as an index of measurement reliability. In Methods of clinical epidemiology (pp. 39-49). Berlin, Heidelberg: Springer Berlin Heidelberg.

Narisetty, N. N. (2020). Bayesian model selection for high-dimensional data. In Handbook of statistics (Vol. 43, pp. 207-248). Elsevier.