Poisson regression for asthmatic attacks

MATH-493 - Applied Biostatistics: Individual Project

Emanuele Sorgente

2023-05-24

Introduction

In this project, our primary aim is to unravel the possibility of predicting asthmatic attacks. Can we accurately forecast when these attacks will occur? When it comes to estimates the number of asthmatic attacks, Poisson regression is a useful tool. While linear regression may appear to be a practical method for dealing with count data by considering them as continuous numerical variables, it can result in illogical predicted values. For example, using linear regression to estimate the number of asthmatic attacks may yield negative results, which violates clinical intuition. This study's dataset includes information on the number of asthmatic attacks per year among a sample of 120 individuals, as well as the associated factors. The final goal of this project is to investigate the association between the number of asthmatic attacks in 2021 and three sociodemographic parameters. The study utilizes the four variables described below:

- gender: Gender of the subjects (categorical) {female=0, male=1}.
- res_inf: Recurrent respiratory infection (categorical) {no=0, yes=1}.
- ghq12: General Health Questionnare 12 (GHQ-12) score of psychological well being (numerical) {0 to 36}.
- attack: Number of asthmatic attack per year (count).

Data exploration

It is critical to thoroughly study the provided dataset before proceeding with model fitting and tuning. The dataset, as shown in Table 1 of the appendix, consists of 120 rows, each of which corresponds to a different patient. We can see the dependent variable under inquiry as well as the recorded values for the regressors specified in the introduction within each row.

$\mu_{attack} = 2.46;\ \mu_{ghq12} = 16.34;\ N_{male} = 53;\ N_{female} = 67;\ N_{res\_inf\_No} = 51;\ N_{res\_inf\_Yes} = 69.$

$$\sigma^2_{attack} = 4.05;\ \sigma^2_{ghq12} = 96.24.$$

According to the descriptive statistics supplied, the average number of attacks experienced by patients falls between 2 and 3 assaults. In addition, the average ghq12 score is around half of the maximum attainable score of 36. Furthermore, there is a higher proportion of girls than males, as well as a bigger number of people with recurring respiratory infections than those without.Given the close proximity of the assaults' mean and variance, it is acceptable to explore building a Poisson regression model. The Poisson distribution is frequently used to explain count data where the mean and variance are about identical. We may study the link between the predictor variables and the frequency of attacks using a Poisson regression model that accounts for covariate effects.

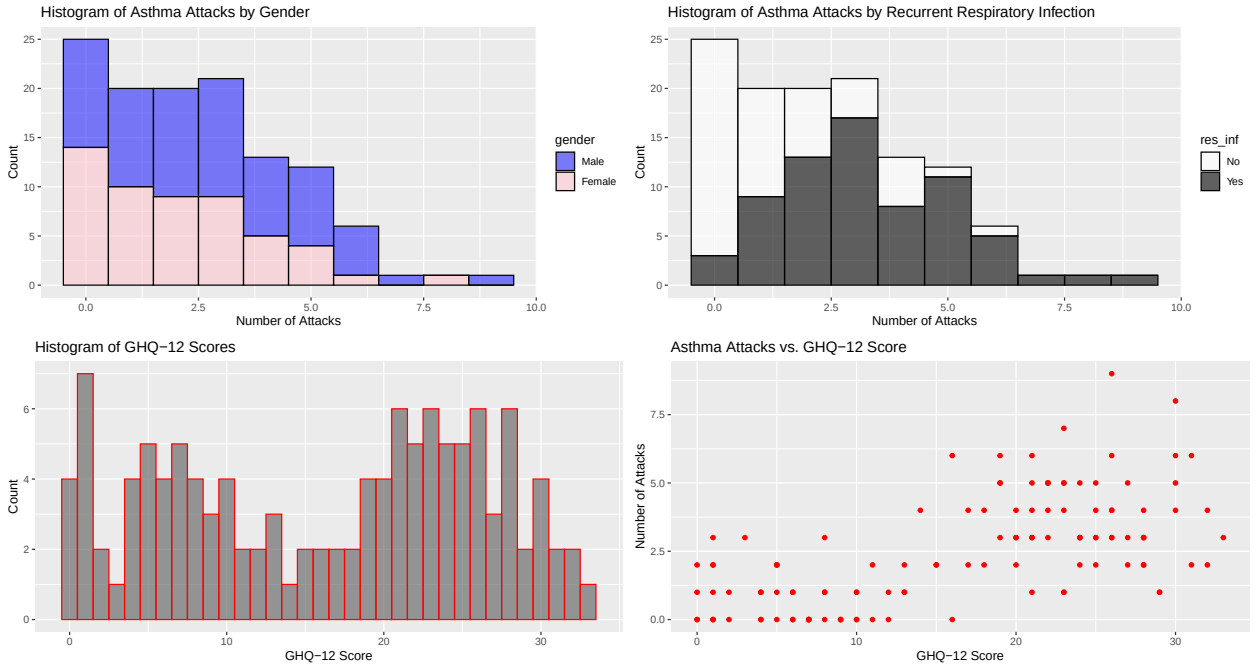

Figure 1: Figure 1: Relevant plots used for data exploration

The findings depicted in Figure 1 demonstrate several noteworthy observations. In the first plot, it is evident that the mode of the distribution for the number of asthma attacks is zero. However, a substantial proportion of individuals experience 1 to 3 attacks, with a notable prevalence of 3 attacks compared to 1 or 2. Notably, among those with zero asthma attacks, the number of females exceeds that of males. Conversely, as the number of attacks increases, a discernible majority of males emerges. Moving to the second plot, it becomes apparent that a significant portion of individuals without asthma attacks also do not exhibit recurrent respiratory infections. In contrast, the average number of asthma attacks is considerably higher among patients with a history of recurrent respiratory infections than those without such infections. Examining the third plot reveals a distinct distribution pattern for the GHQ-12 score, characterized by three distinct clusters. The first cluster is centered around scores of 0 and 1, the second cluster ranges between 5 and 10, and the largest cluster is situated approximately between scores of 20 and 30. This distribution pattern resembles a Gaussian mixture model with three clusters, with the mode appearing at a score of 1. Finally, the

last plot illustrates a positive correlation between the number of asthma attacks and the GHQ-12 score, indicating a trend where an increased number of attacks is associated with higher scores on the GHQ-12.

Poisson regression

Poisson regression is a generalized linear model form of regression analysis to model count data and contingency tables. Poisson regression assumes that the response variable Y follows a Poisson distribution and that the logarithm of its expected value can be modeled by a linear combination of the explanatory variables. Poisson regression models are generalized linear models with the logarithm as the link function and the Poisson distribution as the assumed response probability distribution. If $X \in \mathbb{R}^p$ is a vector of explanatory variables (covariates), then the model takes the form

$$\log(\mathbb{E}(Y \mid X)) = \beta_0 + \beta^T X \tag{1}$$

where $\beta_0 \in \mathbb{R}$ and $\beta \in \mathbb{R}^p$. This can be written more compactly as

$$\log(\mathbb{E}(Y \mid X)) = \beta^T X \tag{2}$$

where X is now a $(p+1)$-dimensional vector consisting of p covariates concatenated with number one, and β is now concatenated with β_0.

If y_i are independent observations with corresponding values x_i of the predictor variables, then β is estimated by maximum likelihood.

The log-likelihood function is derived based on the Poisson distribution and the assumed relationship between the mean and predictors. The log-likelihood function for the GLM with a Poisson distribution is given by

$$L(\beta) = \sum_{i=1}^{n} (y_i \log(\mu_i) - \mu_i) \tag{3}$$

where:

- $\mu = \log(\mathbb{E}(Y \mid X))$
- n is the number of observations.
- y_i is the observed response variable for the ith observation.

The log-likelihood function is maximized to estimate the parameters. This maximization is done using an iterative numerical optimization algorithms, the Newton-Raphson scoring method. The algorithm iteratively updates the parameter estimates until convergence is reached.

A characteristic of the Poisson distribution is that its mean is equal to its variance. In many circumstances it is found that the observed variance is greater than the mean; this is known as overdispersion and indicates that the model is not appropriate. In this case a less restrictive model called Negative Binomial regression is more appropriate.

Model fitting

Base model.

We assume that the observations of target and explanatory variables are i.i.d. variables drawn from a joint distribution (X, Y)

$$(y_1, x_1), \ldots, (y_n, x_n) \sim (X, Y). \tag{4}$$

Initially, we employed a straightforward regression model that incorporated all variables without any interactions. This model served as our "base model" from which we commenced our investigation to glean insights regarding the subsequent process of model selection. Let

$$Y = attack;\ X_1 = gender;\ X_2 = res\_inf;\ X_3 = ghq12.$$

Then, we fit the following model:

$$\log(\mathbb{E}(\hat{Y} \mid X)) = \beta_0 + \beta_{\mathbf{1}} X_1 + \beta_{\mathbf{2}} X_2 + \beta_{\mathbf{3}} X_3 \tag{5}$$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.3153873	0.1834997	-1.7187346	0.0856627
gendermale	-0.0419050	0.1224687	-0.3421692	0.7322236
res_infyes	0.4264306	0.1528588	2.7897024	0.0052757
ghq12	0.0495082	0.0078777	6.2846247	0.0000000

Based on the output analysis, we observed that the variable of gender did not exhibit statistical significance, indicated by a p-value greater than 0.05. Consequently, we fit a revised model, omitting the gender variable. We have now the following model:

$$\log(\mathbb{E}(\hat{Y} \mid X)) = \beta_0 + \beta_{\mathbf{1}} X_2 + \beta_{\mathbf{2}} X_3 \tag{6}$$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.3405072	0.1682250	-2.024117	0.0429581
res_infyes	0.4281601	0.1528235	2.801664	0.0050840
ghq12	0.0498870	0.0077895	6.404364	0.0000000

Based on the analysis of the output, we determine that both variables are statistically significant predictors of asthmatic attacks. This preliminary model forms the basis for further investigation and analysis. The value of Akaike Information Criterion for this model is $AIC = 417.75$

Alternative model.

Subsequently, we fit an alternative model to explore the potential improvement achieved by introducing an interaction term between the two variables present in the preliminary model. The model is the following:

$$\log(\mathbb{E}(\hat{Y} \mid X)) = \beta_0 + \beta_{\mathbf{1}} X_2 + \beta_{\mathbf{2}} X_3 + \beta_{\mathbf{3}}(X_2 \cdot X_3) \tag{7}$$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.6343596	0.2340798	-2.710014	0.0067280
res_infyes	1.0192691	0.3282231	3.105415	0.0019001
ghq12	0.0683435	0.0118600	5.762511	0.0000000
res_infyes:ghq12	-0.0313528	0.0153090	-2.047995	0.0405605

The statistical analysis has yielded an intriguing finding: the inclusion of the interaction term between the variables "res_inf" and "ghq12" in the model has been determined to be statistically significant, as indicated by a p-value greater than 0.05. Consequently, there is merit in incorporating this interaction term within the final model. However, it is crucial to acknowledge that the inclusion of this interaction term introduces complexities in interpreting individual coefficients. It is noteworthy that the implications of incorporating the interaction term are less pronounced when the primary objective of the modeling exercise is prediction, rather than clinical interpretation of the results. Additionally, an evaluation of the AIC reveals a lower value for the model with the interaction term compared to the model without interactions, with $AIC = 413.74$, so it gives a better performance than the basic poisson model. Nevertheless, given the marginal differences between the model with the interaction term and the simpler model without the interaction term, the basic model is preferred.

Final model. Lastly, our final model is:

$$\log(\mathbb{E}(\hat{Y} \mid X)) = -0.3405 + 0.4281 \cdot res\_inf + 0.0499 \cdot ghq12 \tag{8}$$

Model Assesment

Chi-square goodness of fit. The Chi-square goodness of fit test compares the observed counts with the expected counts based on the Poisson distribution assumption. The p-value of this test represents the probability of obtaining a test statistic as extreme as the observed test statistic, assuming the null hypothesis is true, which in our case is that the observed data follow a Poisson distribution. We then proceed calculating the Chi-square test statistic using the formula:

$$\chi^2 = \sum \frac{(O - E)^2}{E} \tag{9}$$

where:

- O represents the observed counts.

- E represents the expected counts.

We determine the degrees of freedom for the Chi-square test, which is calculated as the number of bins minus 1. Using the Chi-square distribution with the degrees of freedom, we find the cumulative probability of obtaining a test statistic as extreme as or more extreme than the observed test statistic. Finally, we subtract the cumulative probability from 1 to obtain the p-value. A P-value > 0.05 indicates good model fit. In the case of our model is $p = 0.1019$ so we assume our model as well fitted, and the observed counts are consistent with the Poisson assumption.

Independence of the observations The residuals of the model must be examined to assess the independence of the data in our analysis. The residuals are the differences between the observed and predicted values generated by the model. We can visually analyze any patterns or systematic departures that may suggest a lack of independence by graphing the residuals against the index or observation number.

The residuals-versus-index plot, also known as the "Residuals vs Index" plot, allows us to examine the presence of any observable trends, clusters, or serial correlations in the residuals. In an ideal world, we would anticipate the residuals to disperse randomly around zero, with no discernable pattern.

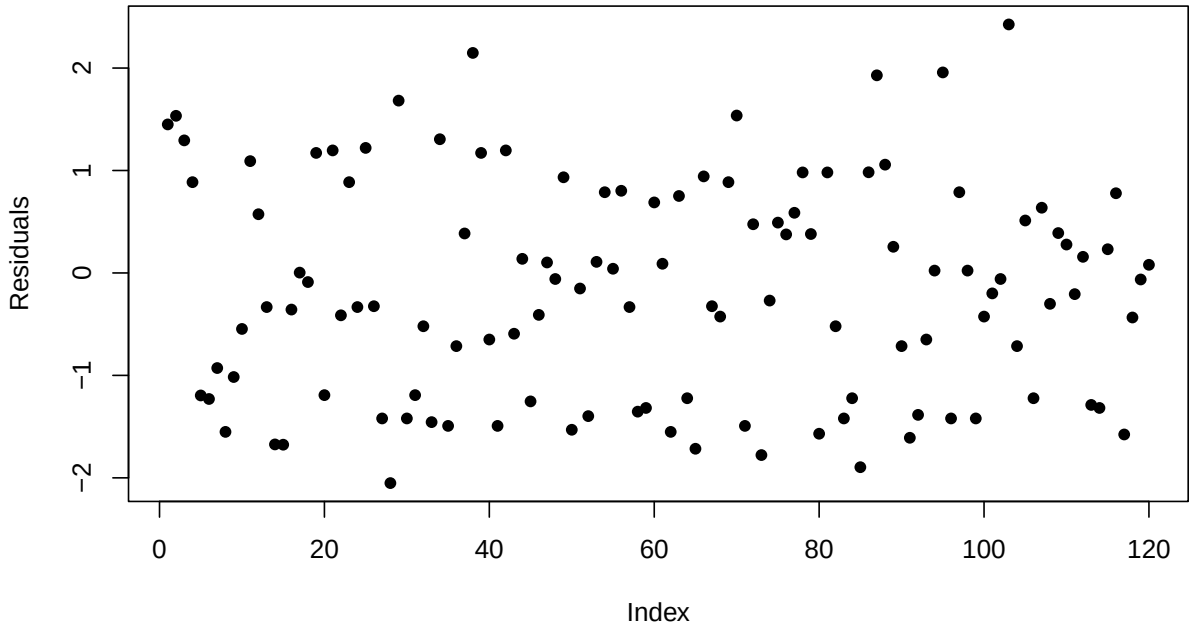

Figure 2: Figure 2: Residuals vs Index

There is no clear pattern or trend found in Figure 2, which depicts the plot of residuals versus index. The figure implies that the residuals have random spread around zero, indicating that

the observations may be independent. We will, however, undertake a Durbin Watson test to further analyze and reinforce our evaluation of independence.

Durbin Watson test The Durbin Watson statistic is a useful test statistic to detect autocorrelation in the residuals from a regression analysis. The formula for the test is:

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

where:

- T is the number of observations.
- e_t are the residuals of the model.

We obtained a test statistic of 1.7784 and a p-value of 0.1107 after running a statistical test on our final model. We reject the null hypothesis because the p-value is greater than the conventional significance level of 0.05. As a result, we conclude that the residuals in our regression model do not exhibit significant autocorrelation.

The absence of autocorrelation in the residuals supports the assumption of independence in our model. This finding supports the validity of our statistical analysis and suggests that the observed residuals are unaffected by the values of previous or subsequent observations.

Overdispersion Overdispersion is a condition in which the variance of the response variable exceeds the mean. A comparison of the deviance residuals and the expected deviance residuals can be made to assess the presence of overdispersion, assuming the null hypothesis that the model is correctly specified. The deviance residuals capture the difference between the observed and expected deviances, whereas the expected deviance residuals are calculated under the assumption of a correctly specified model. If the deviance residuals exceed the expected deviance residuals, overdispersion exists, indicating that the response variable's variance exceeds its mean. This assessment determines whether additional considerations or adjustments are required to account for the data's excess variability.

Deviance residuals are a measure of the model fit, and large or systematic deviations from zero can indicate overdispersion. Plot the residuals of deviance against the predicted values or other relevant variables to identify patterns or trends. To visualize the residuals, use the'residuals' function to extract the deviance residuals from the fitted model and create scatter plots.

According to the analysis of the deviance residuals, there is no discernible pattern or trend with increasing predicted values. This finding indicates that there is no overdispersion in the data. In other words, the response variable's variability is not significantly greater than what would be expected based on the mean. This finding implies that the model's assumptions are correct and that no additional adjustments to account for excess variability are required.

In addition by looking at the predicted mean and predicted variance we observe:

$$\hat{\mu}_{attack} = 2.46;\ \hat{\sigma}^2_{attack} = 1.86$$

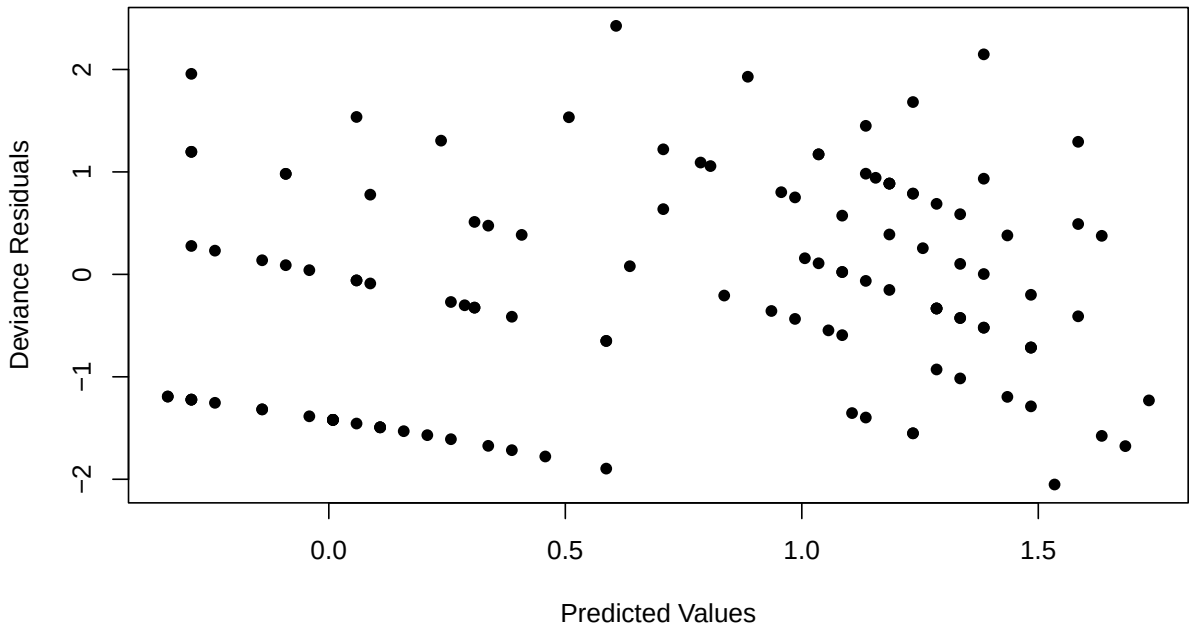

Figure 3: Figure 3: Residuals vs Index

Upon comparing the predicted mean to the actual mean, it is observed that they are approximately equal. However, there is a slight difference in the variances between the predicted and actual data. The slower variance in the predicted data implies that the predicted values are assumed to be less volatile compared to the actual values.

In terms of evaluating the result, a slower variance in the predicted data can be considered favorable in certain contexts. It suggests that the model's predictions are relatively stable and exhibit less volatility compared to the actual data. This can be desirable, especially in situations where consistency and reliability in the predicted outcomes are important.

Outliers We use then Cook's Distance to measure the influence of each observation on the regression coefficients. The Cook's distance statistic is a measure, for each observation in turn, of the extent of change in model estimates when that particular observation is omitted. Any observation for which the Cook's distance is close to 1 or more, or that is substantially larger than other Cook's distances (highly influential data points), requires investigation.

In the plot, Cook's distance values are displayed on the y-axis, representing the influence of each observation. The x-axis represents the observation index or number: As we can see from figure 4 there are no graphical anomalies, there are no observations that we could consider outliers.

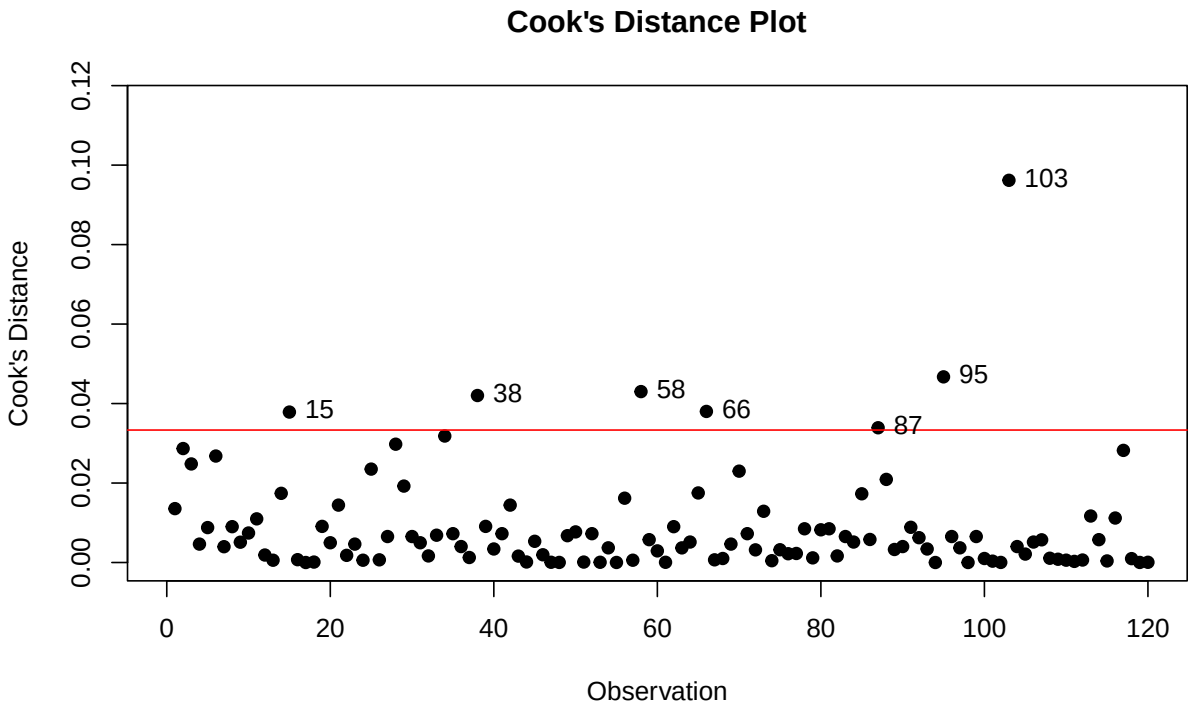

Figure 4: Figure 4: Cook's distance plot

Conclusion

The major goal of this study was to forecast the the amount of occurrences of asthmatic attacks using the available data. We successfully found the suitability of using Poisson regression as a viable strategy for modeling and forecasting count data in a clinical environment through a thorough data exploration process.

Furthermore, we verified the assumptions behind the chosen regression framework. We validated that all key assumptions were upheld by comprehensive review and testing, assuring the validity and reliability of our model.