Generalized Linear Models

Project

Alhamd Faisal - r0767643
Ishika Jain - r0915387
Huu Duc Luu - r0726333
Naichuan Zhang - r0913147
Zarina Serikbulatova - r822631

Group_08

MSc Statistics and Data Science

April 2023

1 Introduction

Falls are a common problem among older adults and can lead to serious injuries, reduced mobility, and even death. The given data has the response variable Y, which records the frequency of falls of subjects of over 65 years of age and in reasonably good health during the six months of the study. The subjects were randomly assigned to the following two interventions: education only (tto = 0) and education plus aerobic exercise training (tto = 1). The other variables which could be considered as important and used as control variables were gender (gender: 0 = female, 1 = male), and balance index (BI). and a strength index (SI). We aim to find whether there is a difference in the number of falls between the subjects having followed the two interventions. We do so by fitting different models, comparing them, and determining the effect of tto based on a better-fitted model.

2 Exploratory data analysis

Exploratory data analysis is performed to better understand the data set given. It is seen that there are two categorical variables, gender(0:Female, 1:Male), tto(0:education only, 1:education with aerobic training). There are 53 males and 47 females in the dataset.

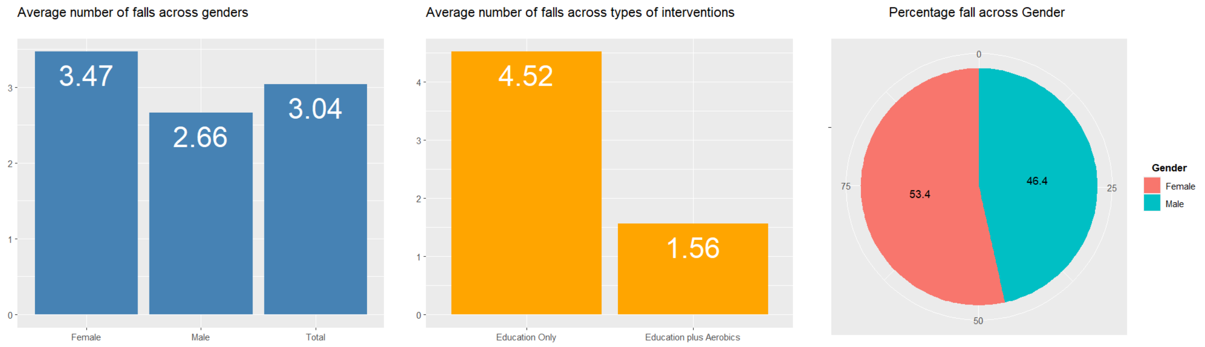

Figure 1: Mean number of falls across genders and interventions.

It can be clearly seen that the mean frequency of falls is higher for the subjects who were given aerobic training than the ones who were not. Also, females on average have a higher number of falls than males. Moreover, the total percentage of falls is also higher in Females compared to Males. Lastly, the score on Balance Index is quite similar in both males and females whereas females on average have a higher score on the strength index than males.

	Balance Index (BI)		Strength Index (SI)	
	Male	Female	Male	Female
Minimum	13	13	29	18
1^{st} **Quartile**	40	37	51	55.50
Median	52	51	58	64
Mean	52.77	52.89	58.62	63.21
3^{rd} **Quartile**	63	71	66	72
Maximum	98	92	87	90

Table 1: Description statistics of balance index and strength index

3 Poisson Model

Poisson model is used to access the effect of two interventions on the frequency of falls during fixed amount of time (6 months study). The model is fitted using *glm* function in R and the results obtained are presented below:

	Estimate	**Standard Error**	**Z-value**	**P-value**	**Risk ratio**	**2.5 %**	**97.5 %**
Intercept	0.489467	0.336869	1.453	0.14623	1.631	0.832	3.114
tto	-1.069403	0.133154	-8.031	9.64e-16	0.343	0.263	0.443
gender	-0.046606	0.119970	-0.388	0.69766	0.954	0.754	1.207
BI	0.009470	0.002953	3.207	0.00134	1.010	1.004	1.015
SI	0.008566	0.004312	0.004312	0.04698	1.009	1.000	1.017

Table 2: Poisson regression estimates and Risk Ratio.

The given model is described by the following equation:

$$Y = exp(-0.490 + (-1.069) * tto + (-0.047) * gender + 0.010 * BI + 0.009 * SI)$$

By taking exponents of coefficients and corresponding confidence intervals for coefficients generates the risk ratios for each of the control variables. The following table provides risk ratios for each of the coefficients. It is visible that while all of the risk ratios except means between the two groups are close to value of 1, while tto(=1) implying people who have undergone both education and aerobic exercises has a value of 0.343. Risk ratio of 0.343 for tto variable indicates that samples who have undergone intervention in the form of education plus aerobic exercise training (tto=1) have 0.343 times average number of falls compared to those who had undergone education only (tto=0). This implies that on average, people who have received education only falls 2.9(=1/0.343) times more than group of people who have gone through both education and aerobic exercises. At a 5% significance level, the confidence interval for tto is [0.263, 0.443]. To check the validity of the model we comparing the null model(with an intercept only) with the fitted model by performing **Likelihood ratio test**.

	LR Chisq.	df	p-value
tto	3.520	1	2.2e-16
gender	0.151	1	0.698
BI	10.282	1	< 0.001
SI	3.976	1	< 0.046

Table 3: Likelihood ratio test

	Test statistic	df	p-value
Pearson Chi-Squared test	105.5466	95	0.2157933
Deviance test	108.7899	95	0.157792

Table 4: Goodness of fit tests

Pearson Chi-Squared test and **Deviance test** were applied to analyze the goodness-of-fit. Pearson Chi-Square tests the hypothesis that the observed counts of the response variable (number of falls) are consistent with the expected value predicted by the Poisson distribution, while the Deviance test is a likelihood-ratio test between the fitted model and the saturated one. As both p-values derived from these tests are greater than the significant level of 0.05, we can conclude that there is no evidence of a significant lack of fit, meaning the Poisson model is a valid model for analyzing the relationship between the dependent and independent variables. To check the validity of the model further, we apply Likelihood Ratio test which compares fitted model with intercept only model.

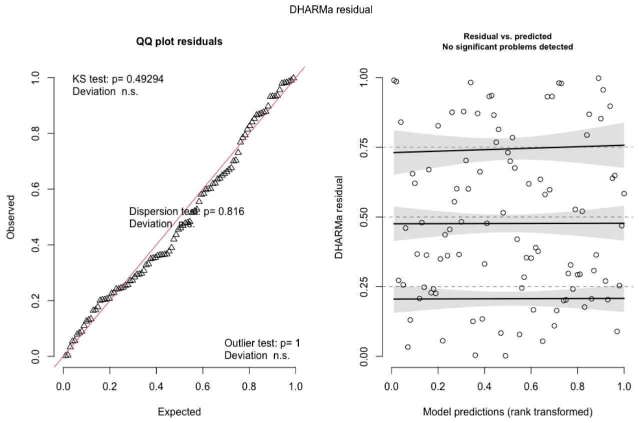

Figure 2: Dharma residual test results.

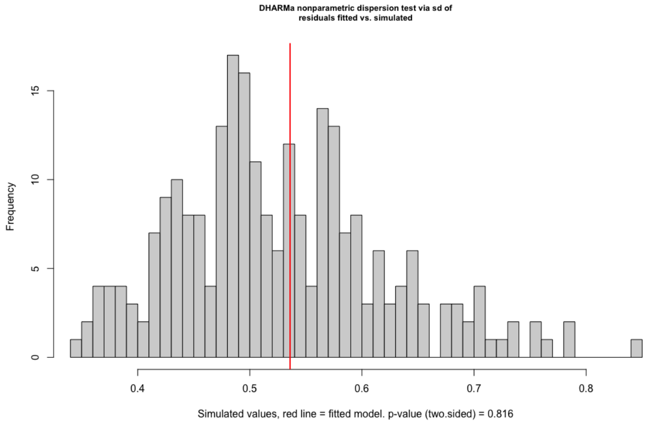

Figure 3: Dharma dispersion test results.

The lack-of-fit was also rejected in the Dharma residuals test by passing the Kolmogorov-Smirnov test, which illustrated that the distribution of the residuals is consistent with the Poisson distribution. With the test statistics of dispersion equals 1.0179 and its p-value of 0.816, there is no evidence of overdispersion as the variance of Y is not very large from the mean.

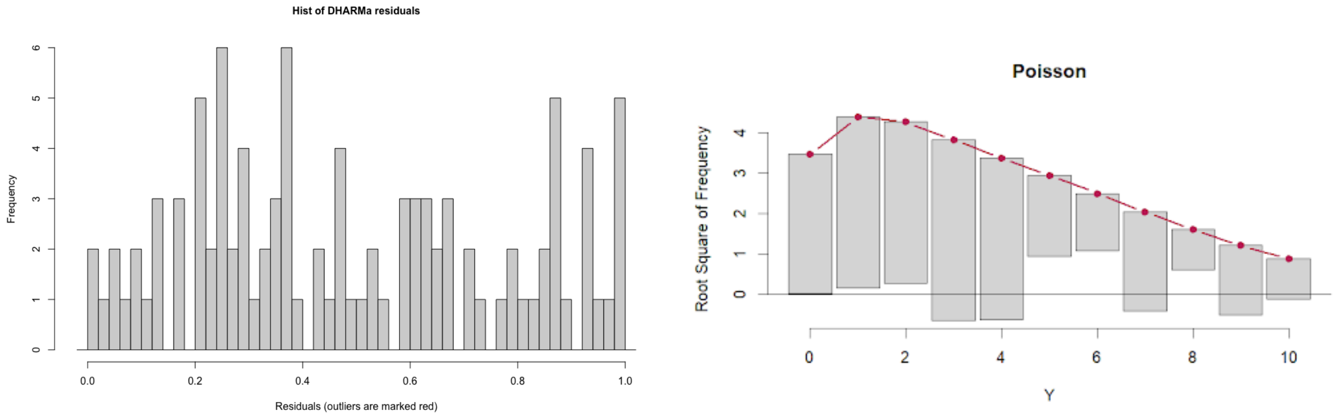

Figure 4: Dharma outlier test results.

In Figure 3, we can not detect any red bars in the left figure, thus no evidence of outliers. Its right side is a rootogram that graphically compares (square roots) of empirical frequencies with fitted frequencies, we can see that there is some over-prediction as well as under-prediction.

4 Quasi-Likelihood Model and Negative Binomial Model

Sometimes classical Poisson model suffers from an overdispersion problem. And then we can use the quasi-likelihood model to assume that the variance is ϕ times the mean. In this case, we have the Poisson overdispersion scale model $var(y_i) = \phi\lambda$. Here we build a quasi-likelihood model and output the results. Also, negative binomial model can solve overdispersion problem. For the negative binomial model, there is a quadratic relationship between the mean and the variance.

$$E(x) = \mu, var(x) = \mu + \frac{1}{\theta}\mu^2$$

In the quasi-likelihood model, $\phi = 1.1110$, which means the variance is estimated 11% larger than the mean. So the problem of overdispersion is not severe. The estimates of quasi-likelihood model and the previous model should be exactly the same, while the estimates of negative binomial model is different but similar. From the estimates we can see that education plus aerobic exercise training can significantly reduce the times that old people falls down. The balance index also has significant effect to avoid falling down.

	Quasi-likelihood	Negative binomial
(Intercept)	0.49	0.49
	(0.36)	(0.34)
tto	-1.07^{***}	-1.07^{***}
	(0.14)	(0.13)
gender	-0.05	-0.05
	(0.13)	(0.12)
BI	0.01^{**}	0.01^{**}
	(0.00)	(0.00)
SI	0.01	0.01^{*}
	(0.00)	(0.00)
AIC		379.29
BIC		394.92
Log Likelihood		-183.64
Deviance	108.79	108.74
Num. obs.	100	100

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 5: Results of quasi-likelihood model and negative binomial model

In the negative binomial model, $\theta = 6060$, which means the quadratic relationship between the mean and the variance is very slight. Here we also draw the rootogram of negative binomial model. We can find that the prediction results are similar to the poisson model. That is we can have a great prediction for people fall 0 2 times but do not have a good prediction for people fall many times. In addtion, our model tends to have a larger predicted value of times of falling. In reality, this results can arise people's awareness to pay attention to falling issue of the elderly people.

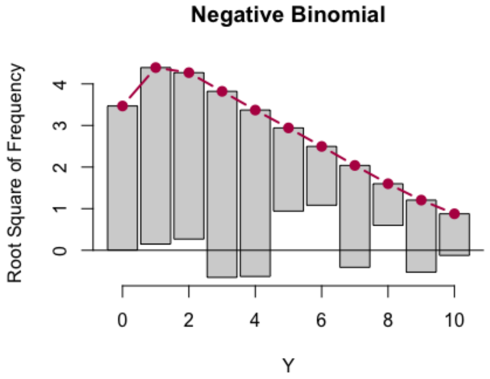

Figure 5: Rootogram of negative binomial model

5 Comparison of Models

The following is the table obtained for the different coefficients and standard errors obtained on fitting the above models:

	Po	**QL**	**NB**	**se.Po**	**se.QL**	**se.NB**
Intercept	0.4895	0.4895	0.48950	0.3369	0.3551	0.3370
tto	-1.0694	-1.0694	-1.0694	0.1332	0.1404	0.1332
gender	-0.0466	-0.0466	-0.0466	0.1200	0.1265	0.1200
BI	0.0095	0.0095	0.0095	0.0030	0.0031	0.0030
SI	0.0086	0.0086	0.0086	0.0043	0.0045	0.0043

Table 6: Model estimates and standard errors.

We see that all the above models led to the same coefficient estimates of predictor variables, and the quasi-likelihood model resulted in standard errors slightly larger than the other two models. We know that while the coefficient estimates for a Poisson and Quasi-Likelihood model will always be identical, the same does not hold for the Negative Binomial model. Finally, on comparing the AIC of the Negative Binomial model, and the Poisson model, we see that the Poisson model performs slightly better than the Negative Binomial model. We cannot calculate an AIC for the Quasi-Likelihood model as there is no likelihood function.

	Df	**AIC**
Poisson	5	377.2878
NB	6	379.2880

Table 7: AIC of Poisson vs. Negative Binomial

6 Conclusion

We conclude that the Poisson model fits the data pretty well, as it does not violate the uniformity assumption and shows no evidence of overdispersion. All the predictor variables apart from gender are significant. The Likelihood ratio and Wald Test give a p-value very close to 0 rejecting the null hypothesis, suggesting the high significance of the predictor variable tto. The goodness of fit tests also does not reject the null hypothesis. Furthermore, the AIC of the Poisson model is lesser than the negative binomial model. Thus we can say that the best-fitted model for the given data is the Poisson model.

According to the Poisson model, the tto or the intervention in which the subjects are passed has a statistically significant impact on the number of falls, which is explainable since the subjects with $tto = 1$ were given aerobic exercise training in addition to the education and aerobic exercise tends to keep your muscles strong, which can help the elderly people maintain mobility, thereby reducing the risk of falling.