Generalized linear model assignment: Group 8

Jack Heller (r0862809) Aleksandra Zdravkovic (r0869484)
Viktoria Kirichenko (r0877202) Medha Hegde (r0872802)
Baris Aksoy (r0869901) Raïsa Carmen (s0204278)

31-12-2021

1 Introduction

This report investigated the link between a persons' race and the number of homicide victims a person knows. 1308 people were asked how many homicide victims they know. The raw data is analysed in section 2 after which several statistical models are explored in section 3. Lastly, section 4 concludes the report.

2 Data exploration

In total, 1308 respondents were asked how many homicide victims they knew. Figure 1 shows the absolute and relative number of respondents for each race that knew 0, 1, 2, 3, 4, 5, or 6 homicide victims. The same summary data is displayed in Table 1. It is clear that there are a lot more white participants in the study (1149 (87.84%) white versus 159 (12.16%) black people were questioned) and the relative frequencies show that black people know more homicide victims on average (0.0922541 known homicide victims per person on average for white and 0.5220126 for black participants).

3 Methodology & results

3.1 Poisson model

Since the number of homocide victims a person knows is count data, a Poisson model is first applied to the data (Table 2).

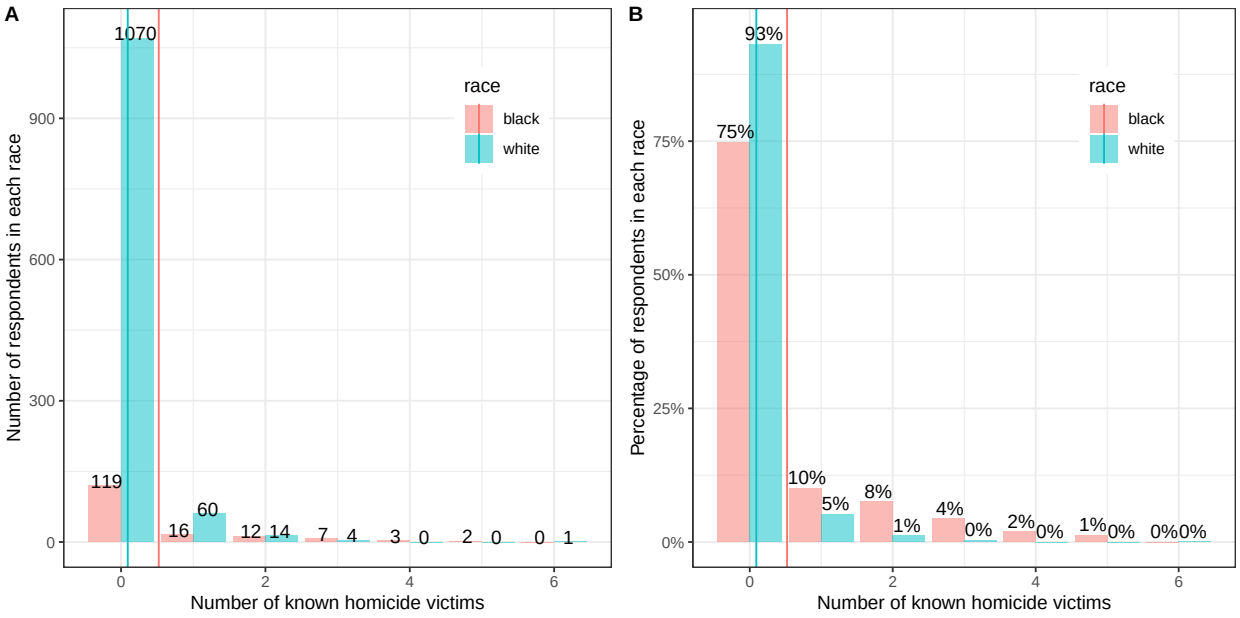


Figure 1: Absolute (A) and relative (B) number of respondents in each race and response group (number of homicide victims the respondent knows). The mean is indicated with a vertical line.

Table 1: Summary data.

Race	Response	Number of respondents	Percentage of respondents withing each race
black	0	119	74.84%
black	1	16	10.06%
black	2	12	7.55%
black	3	7	4.40%
black	4	3	1.89%
black	5	2	1.26%
black	6	0	0.00%
white	0	1070	93.12%
white	1	60	5.22%
white	2	14	1.22%
white	3	4	0.35%
white	4	0	0.00%
white	5	0	0.00%
white	6	1	0.09%

Table 2: Poisson model.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.6500636	0.1097642	-5.922366	0
racewhite	-1.7331446	0.1465678	-11.824866	0

Table 3: Poisson risk ratios.

	RR	2.5 %	97.5 %
(Intercept)	0.52	0.42	0.64
racewhite	0.18	0.13	0.24

The model shows that white respondents know less homicide victims, on average, than black respondents. Indeed, the risk ratio in table 3 shows that the number of known homicide victims for white respondents is 0.18 (between 0.13 and 0.24 with a confidence level of 95%) times the average number of homicide victims that black respondents know on average. Since the poisson regression models the log mean of the poisson regression, the mean number of homicide victims for black people is estimated to be exp(-0.6500636) = 0.5220126 and exp(-0.6500636 + -1.7331446) = 0.0922541 for white individuals. Those averages are exactly equal to the observed values (section @(ref:EDA)). The ratio of the mean responses is 5.6584194 (black/white) and 0.1767278 (white/black). This means that, on average, a black person knows 5.66 times more homicide victims than a white person notice that 0.1767278 = 1/5.66.

Figure 2 compares the true data with the predicted probabilities. Although the model is very accurate with respect to the mean, it is clear that the variance is larger in reality than in the Poisson model. Furthermore, there may be some zero-inflation, especially for the black population.

Indeed, one important assumption in a Poisson model, is that the mean is equal to the variance. The variance in the data is 0.2950959 (1.1498288 for black and 0.1552448 for white respondents). Figure **??** below shows there is overdispersion (the real variance in red is larger than the simulated variance).

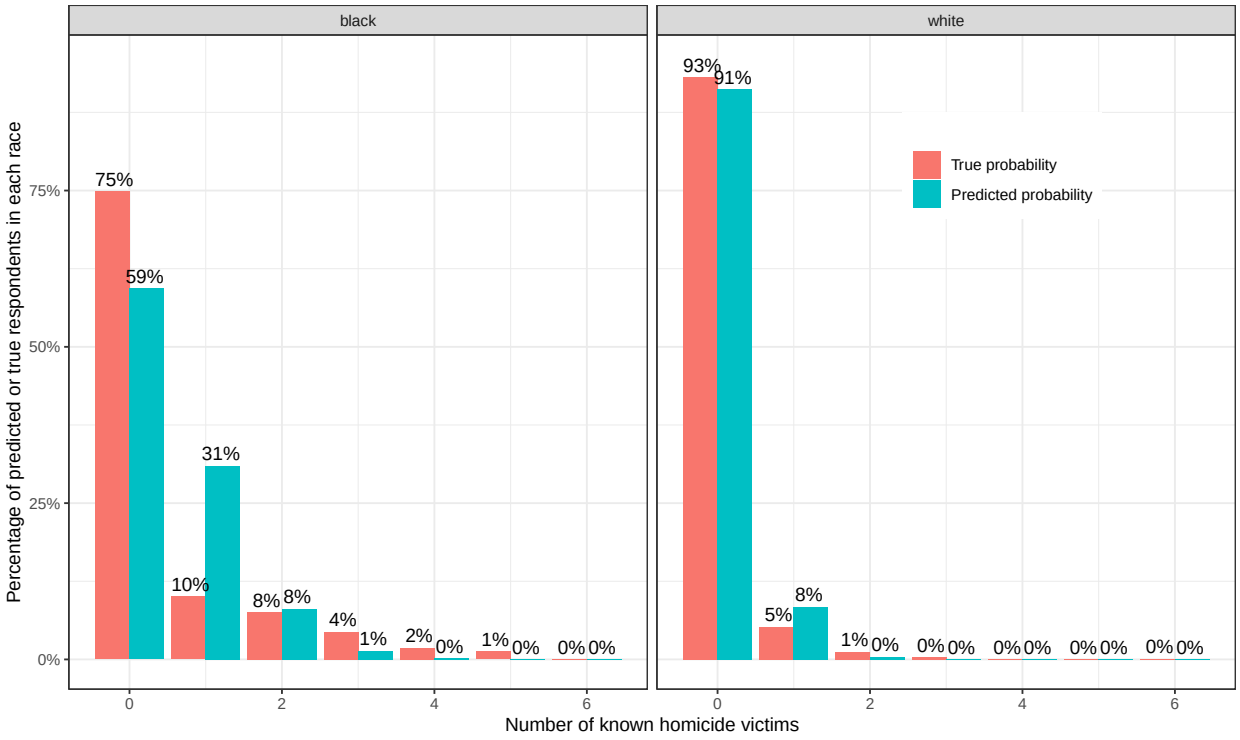
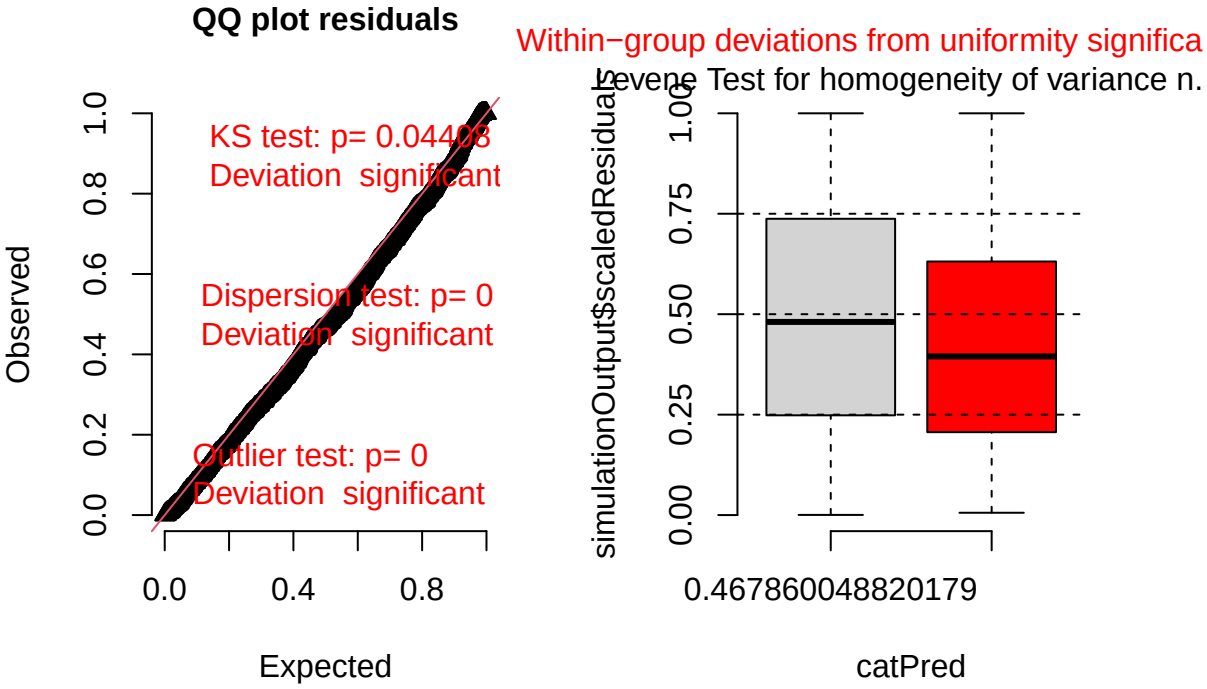
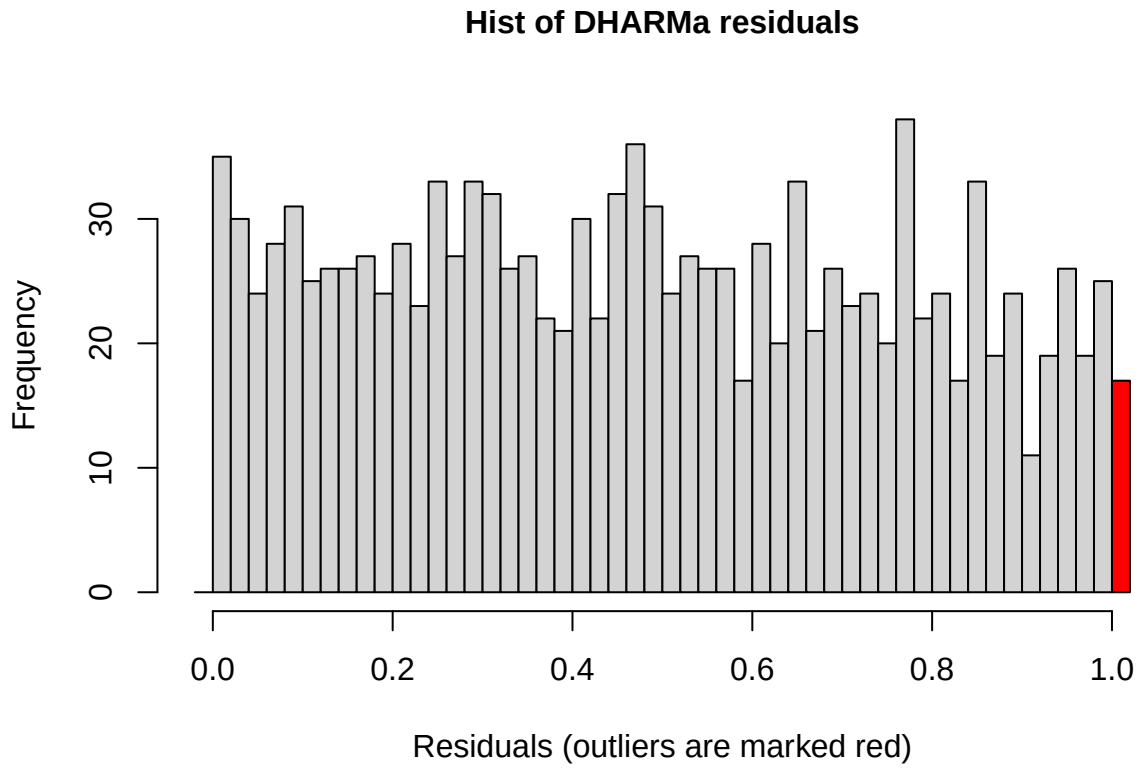
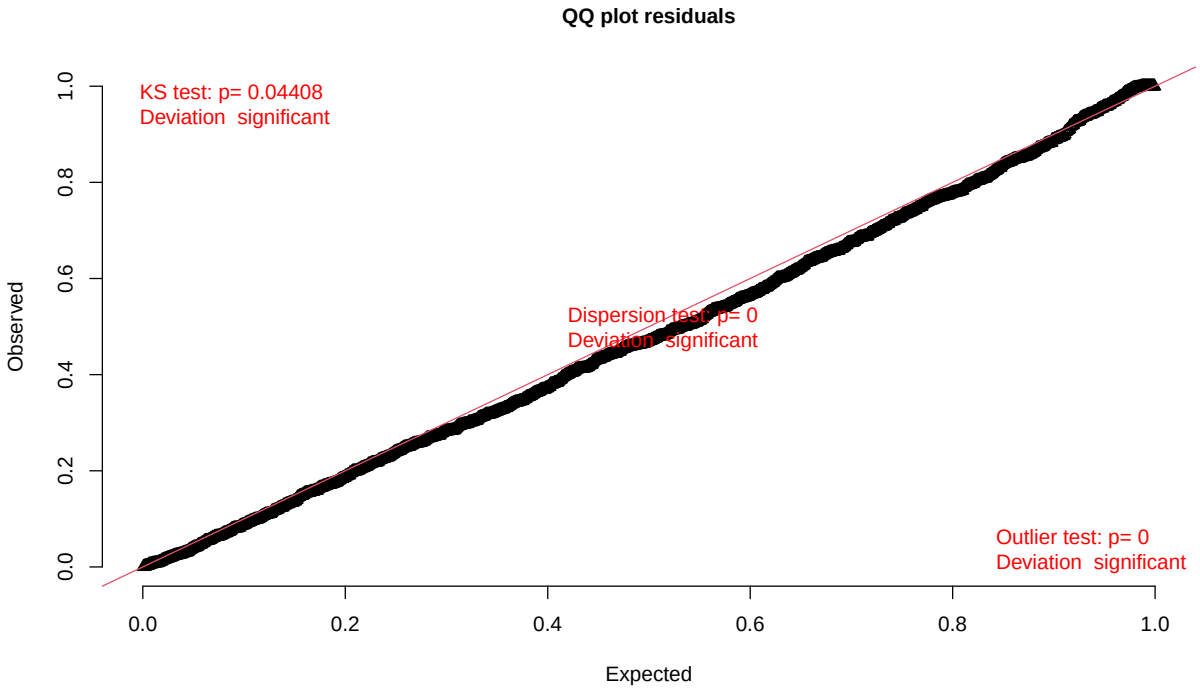


Figure 2: True and predited probabilities of respondents knowing a certain number of homicide victims, for each race.

DHARMa residual








```
##
##  One-sample Kolmogorov-Smirnov test
```

```
##
## data:  simulationOutput$scaledResiduals
## D = 0.038187, p-value = 0.04408
## alternative hypothesis: two-sided
```

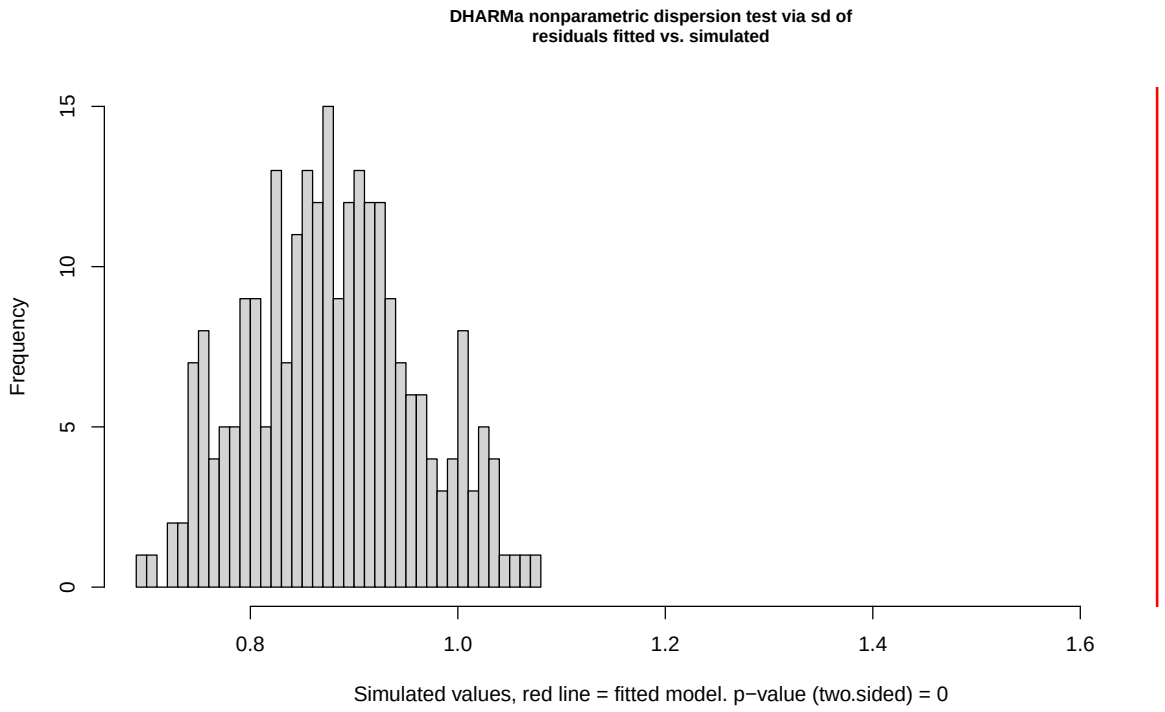


Figure 3: Test of uniformity (A) and dispersion (B)

```
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 1.9031, p-value < 2.2e-16
## alternative hypothesis: two.sided
```

3.2 Negative-binomial model

This model uses the Pascal distribution which counts the number of failures before the yth success. If $x \sim NB(y, \pi)$ with π the probability of success;

$$E(x) = \mu = \frac{y\pi}{1-\pi}$$
$$Var(x) = \sigma^2 = \frac{y\pi}{(1-\pi)^2} = \mu + \frac{1}{\theta}\mu^2 \tag{1}$$

This means that the negative binomial model assumes a quadratic relationship between the mean and the variance.

The variance for each of the races can be obtained from the equation $\sigma^2 = \mu + \frac{1}{\theta}\mu^2$ where $\mu = e^{x'\hat{\beta}}$. For black people, $\mu = 1.9156627$ and the variance is 20.0547984. For white people, $\mu = 0.3385508$ and the variance is 0.9050853.

3.3 Quasi-likelihood model

In quasi-likelihood models, the mean and variance function are specified separately. This thus lifts the poisson assumption that mean and variance are equal. In general, if the mean structure is specified as $\lambda = \mu(x,\beta) = e^{x'\beta}$, then the variance is $var(y_i) = \phi\lambda$ where $\hat{\beta}$ and ϕ are estimated from the Pearson statistic. This model thus assumes a linear relationship between the mean and variance.

```
##
## Call:
## glm(formula = resp ~ race, family = quasipoisson, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0218  -0.4295  -0.4295  -0.4295   6.1874
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.6501     0.1450  -4.482 8.03e-06 ***
## racewhite    -1.7331     0.1937  -8.950  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 1.745694)
##
##     Null deviance: 962.80  on 1307  degrees of freedom
## Residual deviance: 844.71  on 1306  degrees of freedom
## AIC: NA
```

```
##
## Number of Fisher Scoring iterations: 6
```

The regression results show that the dispersion parameter ψ is estimated to be 1.745694 which is much larger than one (Poisson model assumes it to be one).

3.4 Sandwich-estimator

```
## Warning: package 'sandwich' was built under R version 4.1.2
```

```
## Warning: package 'lmtest' was built under R version 4.1.2
```

```
##
## z test of coefficients:
##
##             Estimate Std. Error z value  Pr(>|z|)
## (Intercept) -0.65006    0.16239 -4.0030 6.254e-05 ***
## racewhite   -1.73314    0.20551 -8.4335 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4 Conclusion