Generalized linear model assignment: Group 8

Jack Heller (r0862809) Aleksandra Zdravković (r0869484)
Viktoria Kirichenko (r0877202) Medha Hegde (r0872802)
Barış 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 people were asked how many homicide victims they knew. Figure 1 shows the absolute and relative number of people for each race that knew 0, 1, 2, 3, 4, 5, or 6 homicide victims. 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 (0.09 known homicide victims per person on average for white with a variance of 0.16 and 0.52 on average for black participants with a variance of 1.15).

3 Methodology & results

3.1 Poisson model

Since the number of homicide victims a person knows is count data, a Poisson model is first applied to the data (Table 2 in the appendix). The model shows that, on average, white people know less homicide victims than black people. Risk ratios in table 1 show that the number of homicide victims known by white people is 0.18 times that of black people. To calculate the means for each

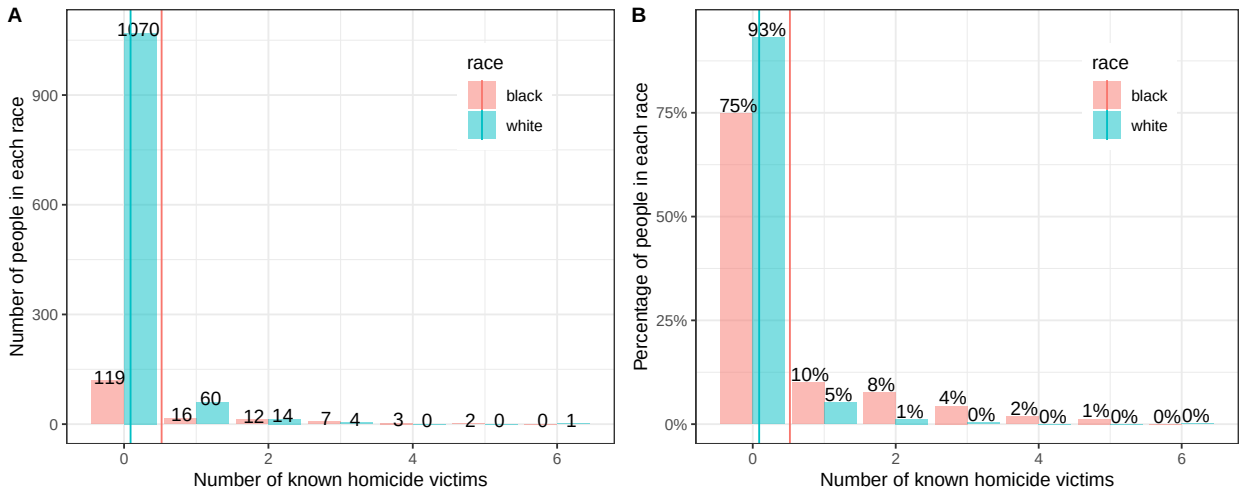


Figure 1: Absolute (A) and relative (B) number of people in each race and response group (number of homicide victims the person knows). The mean is indicated with a vertical line.

Table 1: Poisson risk ratios.

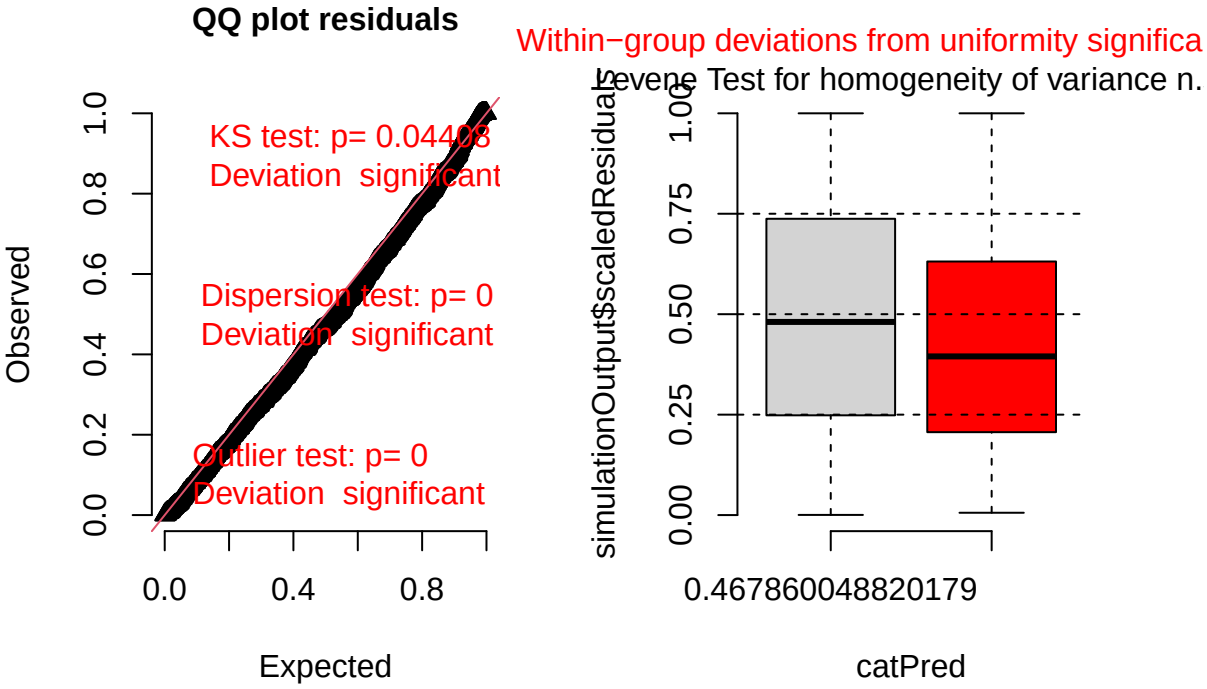
	RR	2.5 %	97.5 %
(Intercept)	0.52	0.42	0.64
racewhite	0.18	0.13	0.24

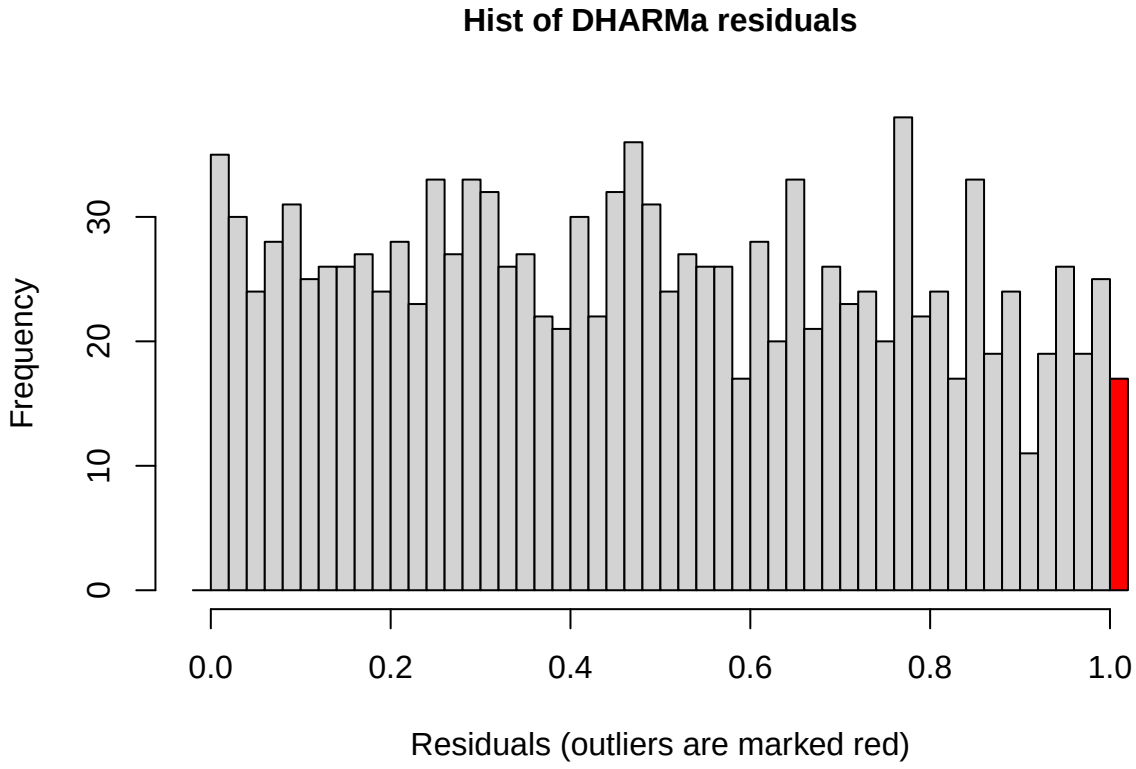
level of covariate, the exponential transformation is being used: $e^{-0.65} = 0.52$ and $e^{-0.65-1.73} = 0.09$ for white individuals. The ratio of the mean responses is 5.66 (black/white) and 0.18 (white/black). Meaning that, on average, a black person knows 5.66 times more homicide victims than a white person.

Figure 5 in the appendix compares the true data with the predicted probabilities. Although the model is very accurate with respect to the mean, the variance is larger in reality than in the Poisson model. Furthermore, there may be some zero-inflation, especially for the black population.

An important assumption in a Poisson model is that mean and variance are equal. The variance in the data is 0.295 (1.15 for black and 0.16 for white people), while the mean is 0.14. Indeed, Figure **??** shows there is overdispersion (the real variance in red is larger than the simulated variance).

DHARMa residual





Kolmogorov-Smirnov test in the residuals' QQ plot indicates a borderline uniformity violation. However, when performing visual inspection of the histogram of DHARMa residuals, the bars do not appear to follow a uniform distribution.

The rootogram shows that level 1 is overpredicted by our model, while other levels are moderately underpredicted.

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  simulationOutput$scaledResiduals
## D = 0.038187, p-value = 0.04408
## alternative hypothesis: two-sided

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

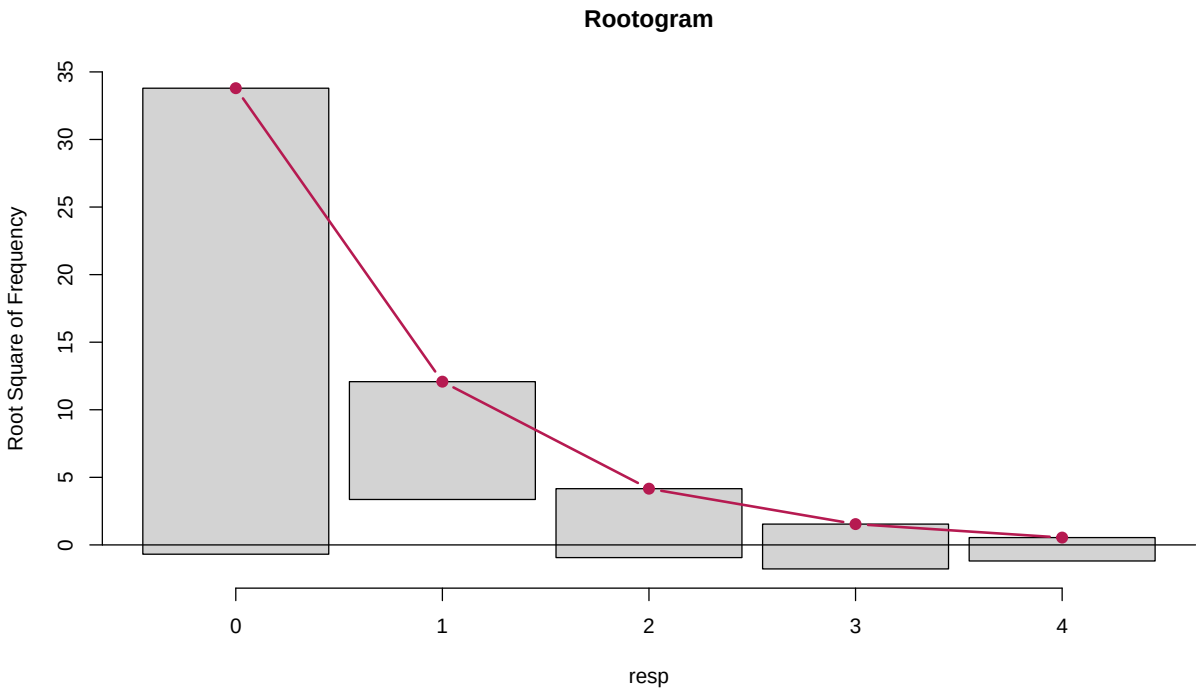


Figure 2: Rootogram

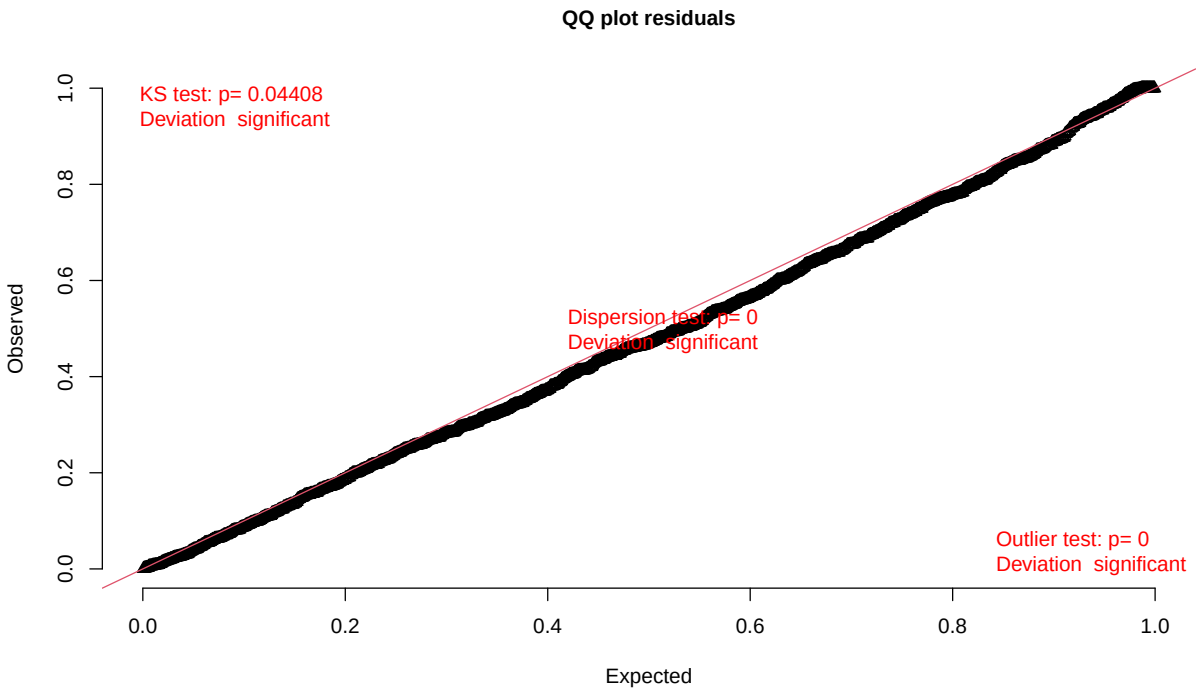


Figure 3: Test of uniformity and dispersion

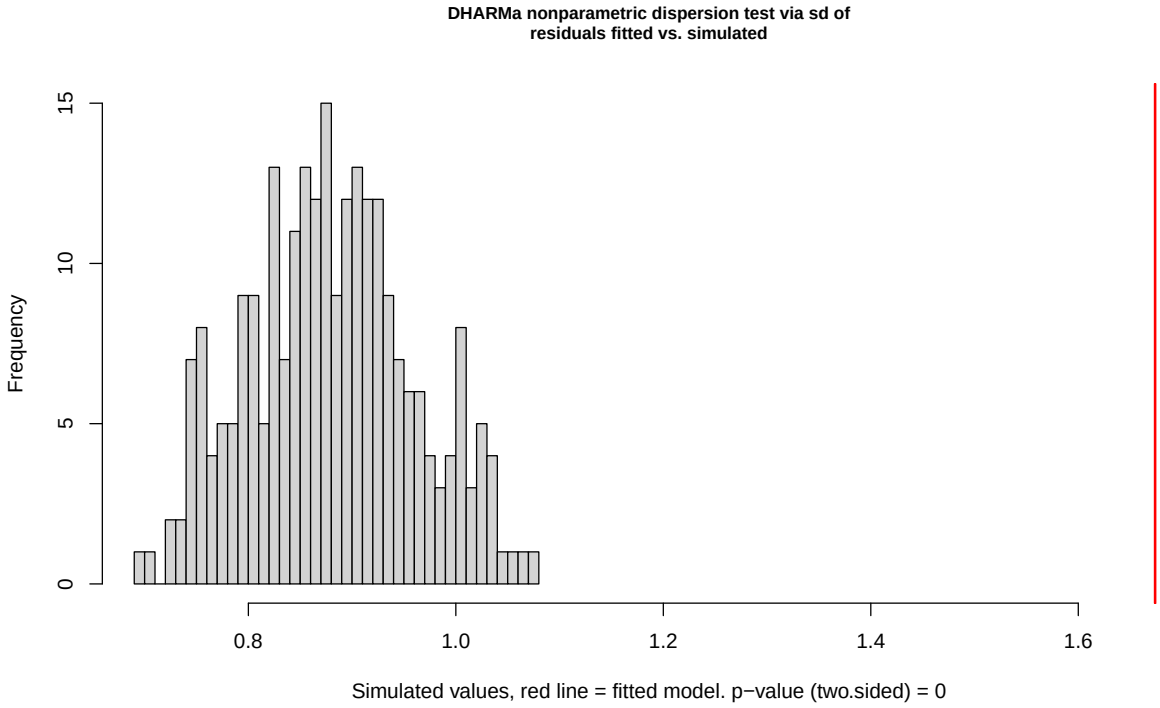


Figure 4: Test of uniformity and dispersion

```
## alternative hypothesis: two.sided
```

The red line, which represents the observed variance, is significantly larger than what we would expect under the model. This indicates the presence of overdispersion.

3.2 Negative-binomial model

The negative binomial model assumes a quadratic relationship between the mean and the variance. The model shows similar estimated coefficients but much larger standard deviations than the Poisson model (table 2). 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, mean $\mu = 0.52$ and the variance is 1.86. For white people, $\mu = 0.09$ and the variance is 0.13. This shows that the variance for black people is overestimated (it was 1.15 in reality) and the variance for white people is slightly underestimated (it was 0.16 in reality).

3.3 Quasi-likelihood model

Quasi-likelihood model lift 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.

Regression results (table 2) show that the dispersion parameter ϕ is estimated to be 1.75, meaning the variance is estimated to be 75% larger than the mean (Poisson model assumes mean and variance are equal). The variances are estimated to be 0.91 for black people and 0.16 for white people.

3.4 Zero-inflated models

Lastly, a zero-inflated Poisson model and Negative Binomial was tested because the raw data showed that there were many people that knew no homicide victims.

The zero-inflated poisson (zip) model shows that white people are significantly more likely to know no homicide victims and the poisson regression coefficient for white people is still negative and highly significant (table 2). The mean and variance are calculated as $\mu_i = (1-\pi_i)\lambda_i$ and $\sigma_i^2 = \mu_i + \frac{\pi_i}{(1-\pi_i)}(\mu_i^2)$ with $i \in \{white, black\}$ where λ is the average rate in the count process and π the probability of zero. This yields $\mu_{black} = 0.52$ and $\sigma^2_{black} = 1.13$ for black people and $\mu_{white} = 0.09$ and $\sigma^2_{white} = 0.14$ for white people.

For the zero-inflated Negative Binomial (zinb) model, the mean and variance are calculated as $\mu_i = (1-\pi_i)\lambda_i$ and $\sigma_i^2 = (1-\pi_i)\lambda_i(1+\lambda_i(\pi_i+\alpha))$ with $i \in \{white, black\}$ where λ is the average for

Table 2: Comparison of all models with the data. For model parameters, standard deviations are listed between brackets, stars indicate siginifcance (0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1).

	Data	Poisson	Negative binomial	Quasi likelihood	Zero-inflated Poisson	Zero-Inflated Negative binomial
Count						
Intercept		-0.65 (0.11) ***	-0.65 (0.21) **	-0.65 (0.15) ***	0.53 (0.14) ***	0.37 (0.25)
racewhite		-1.73 (0.15) ***	-1.73 (0.15) ***	-1.73 (0.19) ***	-1.00 (0.23) ***	-1.07 (0.28) ***
theta			0.20 (0.04)			3.16 (3.32)
phi				1.75		
Probability of zero						
Intercept					0.81 (0.21) ***	0.57 (0.38)
racewhite					0.94 (0.29) **	0.90 (0.33) **
Estimated mean and variance						
mean black	0.52	0.52	0.52	0.52	0.52	0.52
variance black	1.15	0.52	1.86	0.91	1.13	1.24
mean white	0.09	0.09	0.09	0.09	0.09	0.09
variance white	0.16	0.09	0.13	0.16	0.14	0.14
Model performance						
AIC		1121.99	1001.8		998.74	999

the negative binomial process, π the probability of zero, and $\alpha = \frac{1}{\theta}$ the overdispersion parameter. This yields $\mu_{black} = 0.52$ and $\sigma^2_{black} = 1.24$ for black people and $\mu_{white} = 0.09$ and $\sigma^2_{white} = 0.14$ for white people. It is interesting to see that θ is not significant; a zero-inflated Poisson where (θ = 1) might be just as good.

4 Conclusion

Table 2 shows the different models' estimate coefficients and Akaike Information Criterion (AIC). Estimated mean and variance for both races are also included in the table. All models agree that black people know significantly more homicide victims than white people. The mean and variance for the number of homicide victims a black or white person knows is closest to the sample mean and variance in the zip model and the zinb model is a close second. The zip model also has the lowest AIC. Figure 5 compares the different model graphically. It is clear that the poisson model performs the worst, especially for black people. The negative binomial is slightly better than the poisson model for black people but performs bad for white people. The zero-inflated models are the closest to the true data where zip seems to slightly outperform the zinb model; it almost entirely overlaps with the sample probabilities. Figure5 shows the difference between the predicted and observed probabilities more clearly. Overall, the zip model comes closest to the sample probabilities.

Appendix

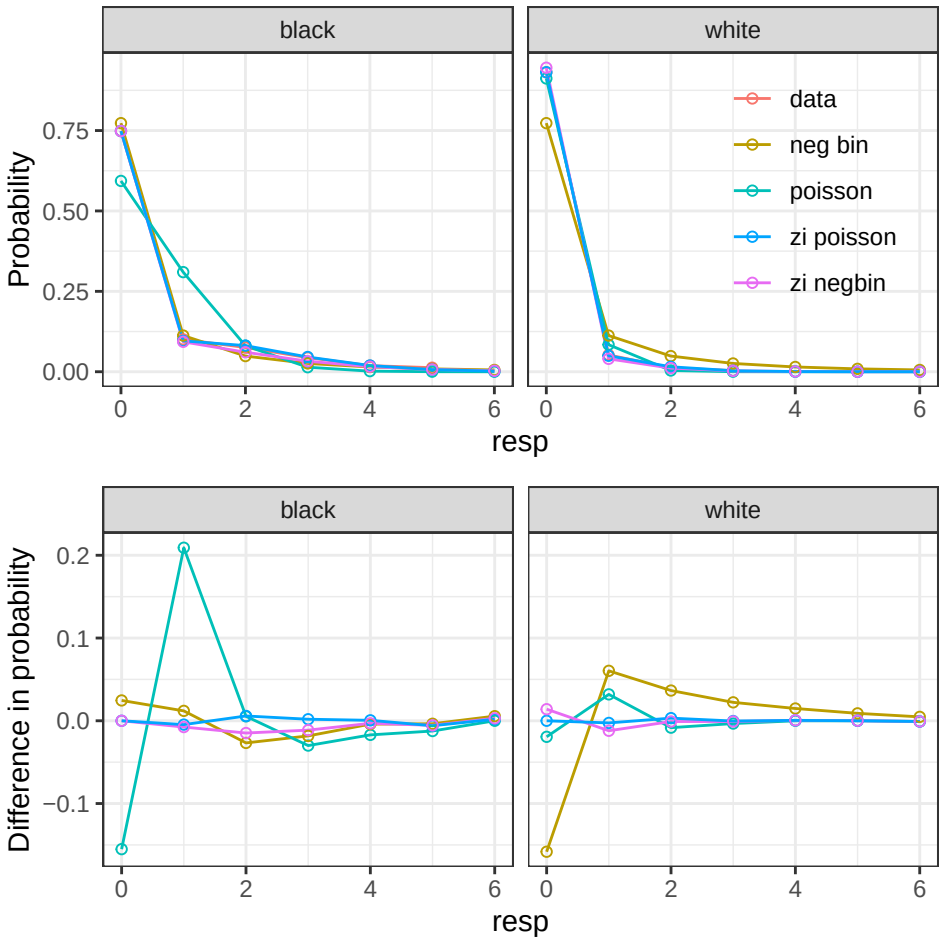


Figure 5: Top figures show the predicted probabilities and sample probabilities in the data. The bottom figures show the predicted probability minus the sample probability.