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 (0.09 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 homocide victims a person knows is count data, a Poisson model is first applied to the data (Table 2).
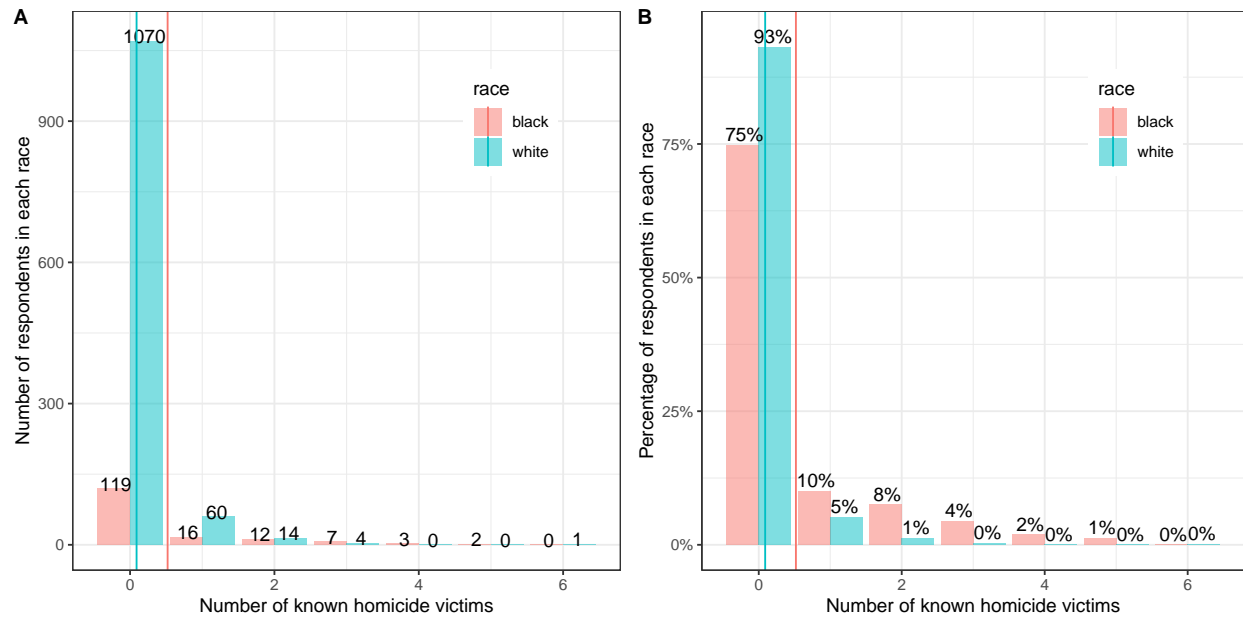
Figure 1: Absolute (A) and relative (B) number of respondents in each race and response group (number of homicide victims the respondent knows). The mean is indicated with a vertical line.

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(-0.65) = 0.52 and exp(-0.65 -1.73) = 0.09 for white individuals. Those averages underestimate the observed values (section 2). The ratio of the mean responses is 5.66 (black/white) and 0.18 (white/black). This means that, on average, a black person knows 5.66 times more homicide victims than a white person notice that 0.18 = 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 (1.15 for black and 0.16 for white respondents). Figure **??** below shows there is overdispersion (the real variance in red is larger than the simulated variance).
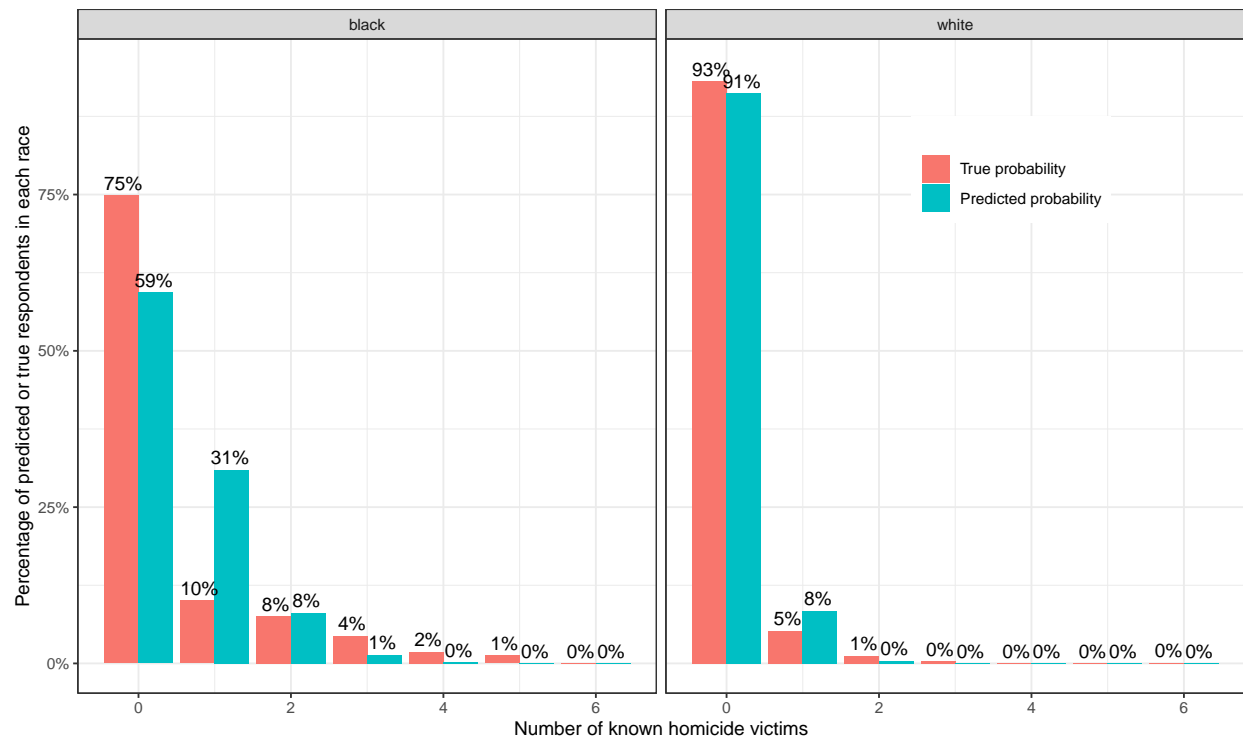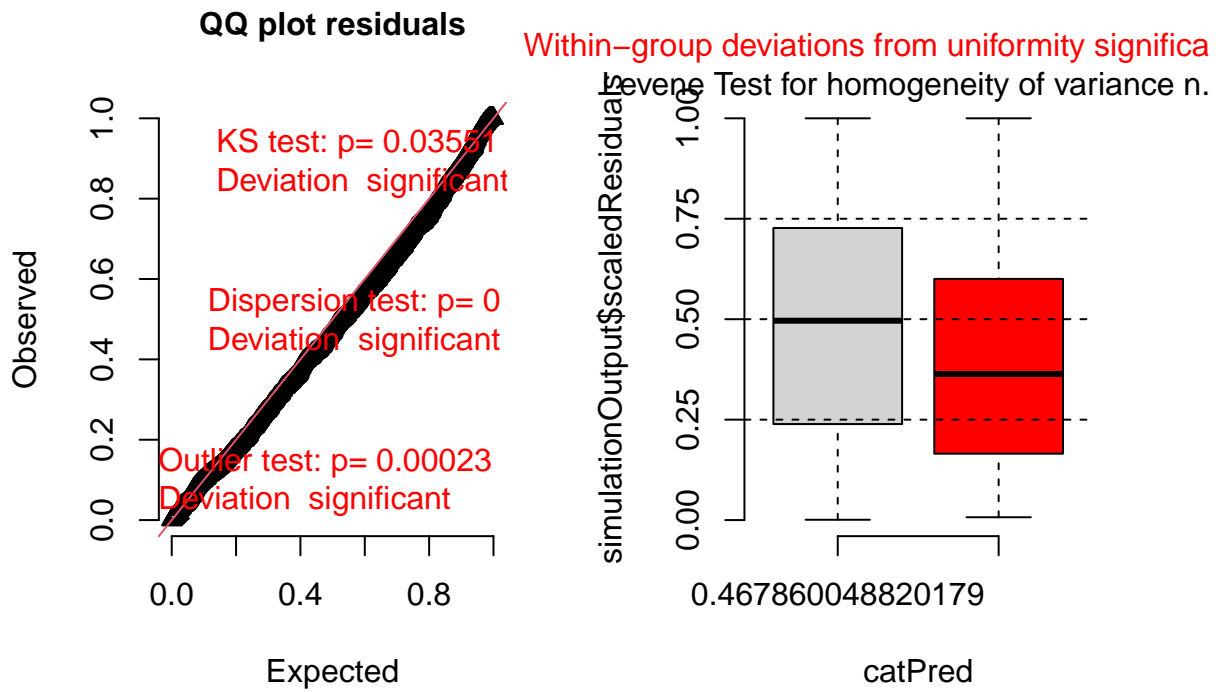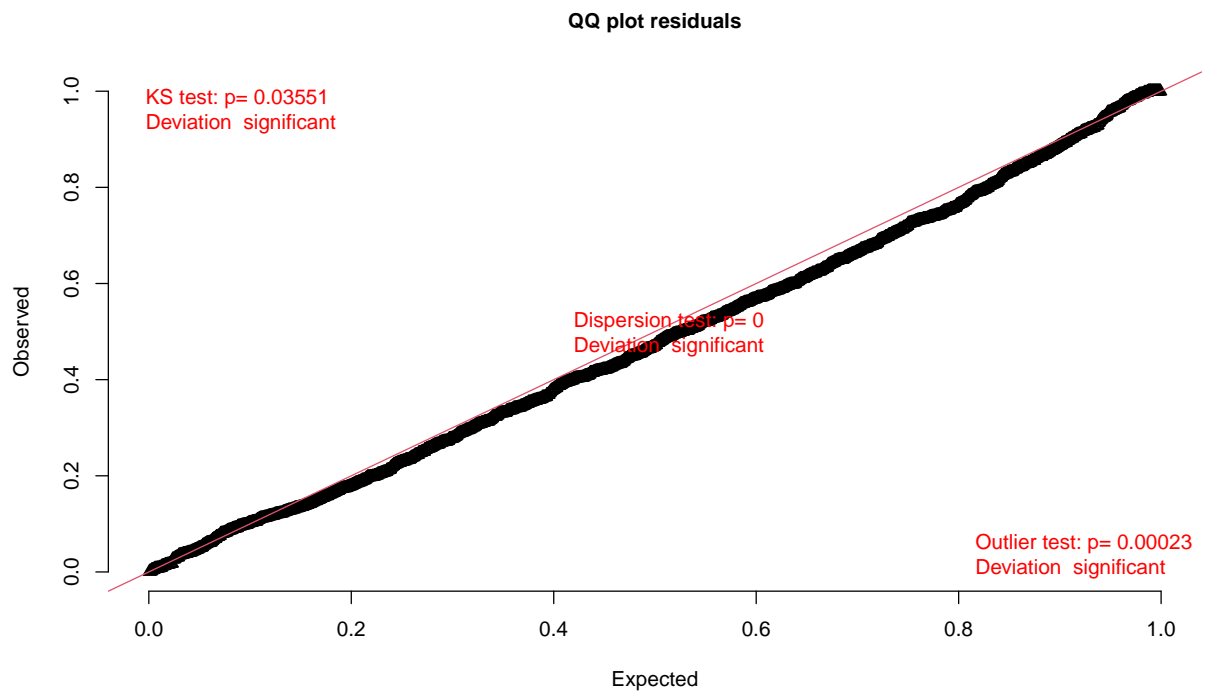
Figure 2: True and predited probabilities of respondents knowing a certain number of homicide victims, for each race.

DHARMa residual

QQ plot residuals

Observed

KS test: p= 0.03551
Deviation significant

Dispersion test: p= 0
Deviation significant

Outlier test: p= 0.00023
Deviation significant

Expected

Within−group deviations from uniformity significa
Levene Test for homogeneity of variance n.

simulationOutput$scaledResiduals

0.467860048820179

catPred

Hist of DHARMa residuals

QQ plot residuals


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

```
##
## data:  simulationOutput$scaledResiduals
## D = 0.039255, p-value = 0.03551
## alternative hypothesis: two-sided
```
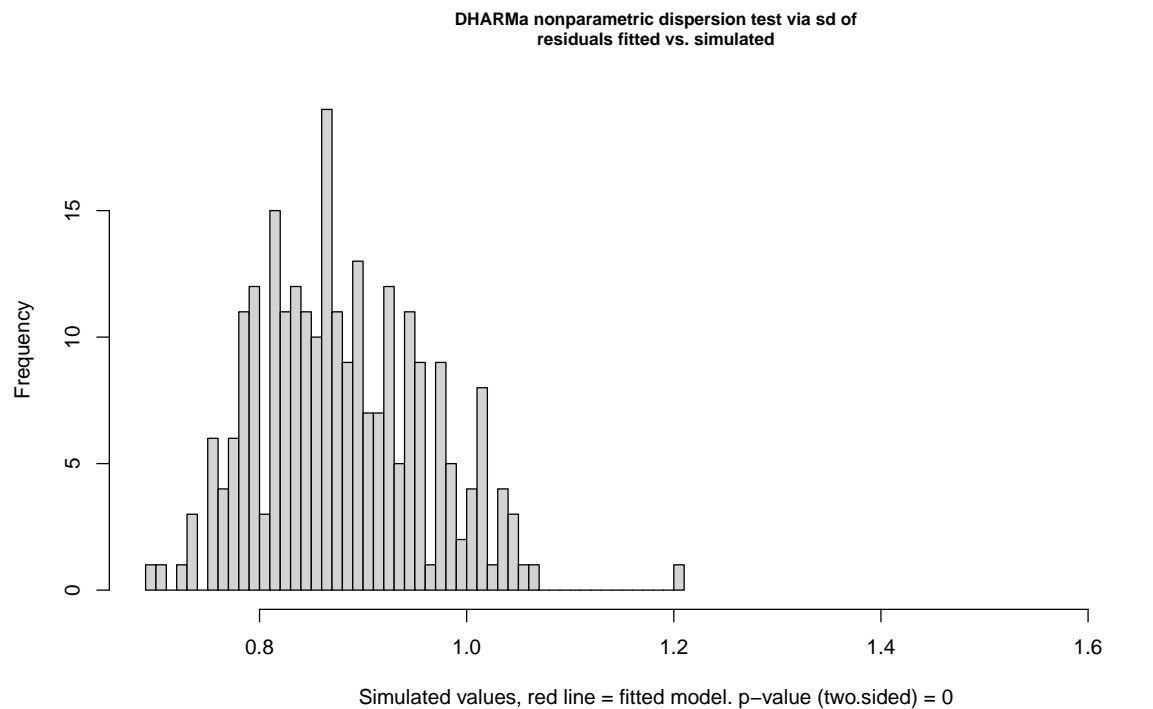

Figure 3: Test of uniformity and dispersion

```
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 1.9005, 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;

Table 4: Negative binomial model.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.6500636	0.2076906	-3.129962	0.0017483
racewhite	-1.7331446	0.2384769	-7.267558	0.0000000
theta	0.2023119	0.0409485		

Table 5: Quasi-likelihood model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.6500636	0.1450256	-4.482406	8e-06
racewhite	-1.7331446	0.1936523	-8.949776	0e+00

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

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

The model shows similar estimated coefficients but much larger standard deviations than the Poisson model (table 4). 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, $\mu = 0.52$ 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

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.

The regression results show that the dispersion parameter ϕ is estimated to be 1.75 which is much larger than one (Poisson model assumes it to be one). The variances are estimated to be 0.91 for

Table 6: Zero-inflated Poisson.

	Estimate	Std. Error	z value	Pr(>|z|)
Count				
(Intercept)	0.5267632	0.1395760	3.774025	0.0001606
racewhite	-1.0049534	0.2310854	-4.348839	0.0000137
Probability of zero				
(Intercept)	0.8083117	0.2122015	3.809171	0.0001394
racewhite	0.9355742	0.2927381	3.195943	0.0013937

black people (it was 1.15in the sample) and 0.16 for white people (1.15in the sample).

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
```

3.5 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 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. 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 $\sigma_{black}^2 = 1.13$ for black people and $\mu_{white} = 0.09$ and $\sigma_{white}^2 = 0.14$ for white people.

Table 7: Zero-inflated Negative Binomial.

	Estimate	Std. Error	z value	Pr(>|z|)
Count				
(Intercept)	0.3667183	0.2544516	1.4412102	0.1495253
racewhite	-1.0728393	0.2752485	-3.8977123	0.0000971
Log(theta)	1.1511039	1.1992330	0.9598668	0.3371223
Probability of zero				
(Intercept)	0.5674072	0.3783452	1.4997076	0.1336902
racewhite	0.9028298	0.3341300	2.7020311	0.0068917

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

4   Conclusion

Table 8 shows the different models' estimate coefficients and Akaike Information Criterion (AIC). The zero-inflated model parameters cannot directly compared to the other three models. Therefore, estimated mean and variance for both races are also included. The mean and variance for the number of homicide victim a black or white person know is closest to the sample mean and variance in the zero-inflated poisson model and the zero-inflated negative binomial is a close second. The zero-inflate Poisson model also has the lowest AIC. Figure @ref(fig: comparemodels2) compares the different model graphically. It is clear that the poisson model performs the worst, especially for black people. The zero-inflated models are the closest to the true data where zero-inflated Poisson seems to slightly outperform the zero-inflated negative binomial model; it almost entirely overlaps with the sample probabilities.

Warning: Removed 3 rows containing missing values (geom_point).

Table 8: Comparison of all models.

|	Poisson	Negative binomial	Quasi likelihood	Zero-inflated Poisson	Zero-Inflated Negative binomial
Count					
Intercept	-0.65	-0.65	-0.65	0.53	0.37
Intercept std	0.11	0.21	0.15	0.14	0.25
racewhite	-1.73	-1.73	-1.73	-1.00	-1.07
racewhite std	0.15	0.15	0.19	0.23	0.28
Probability of zero					
Intercept				0.81	0.57
Intercept std				0.21	0.38
racewhite				0.94	0.90
racewhite std				0.29	0.33
Estimated mean and variance					
mean black	0.52	0.52	0.52	0.52	0.52
variance black	0.52	1.86	0.91	1.13	1.24
mean white	0.09	0.09	0.09	0.09	0.09
variance white	0.09	0.13	0.16	0.14	0.14
Model performance					
AIC	1121.99	1001.80		998.74	999.00
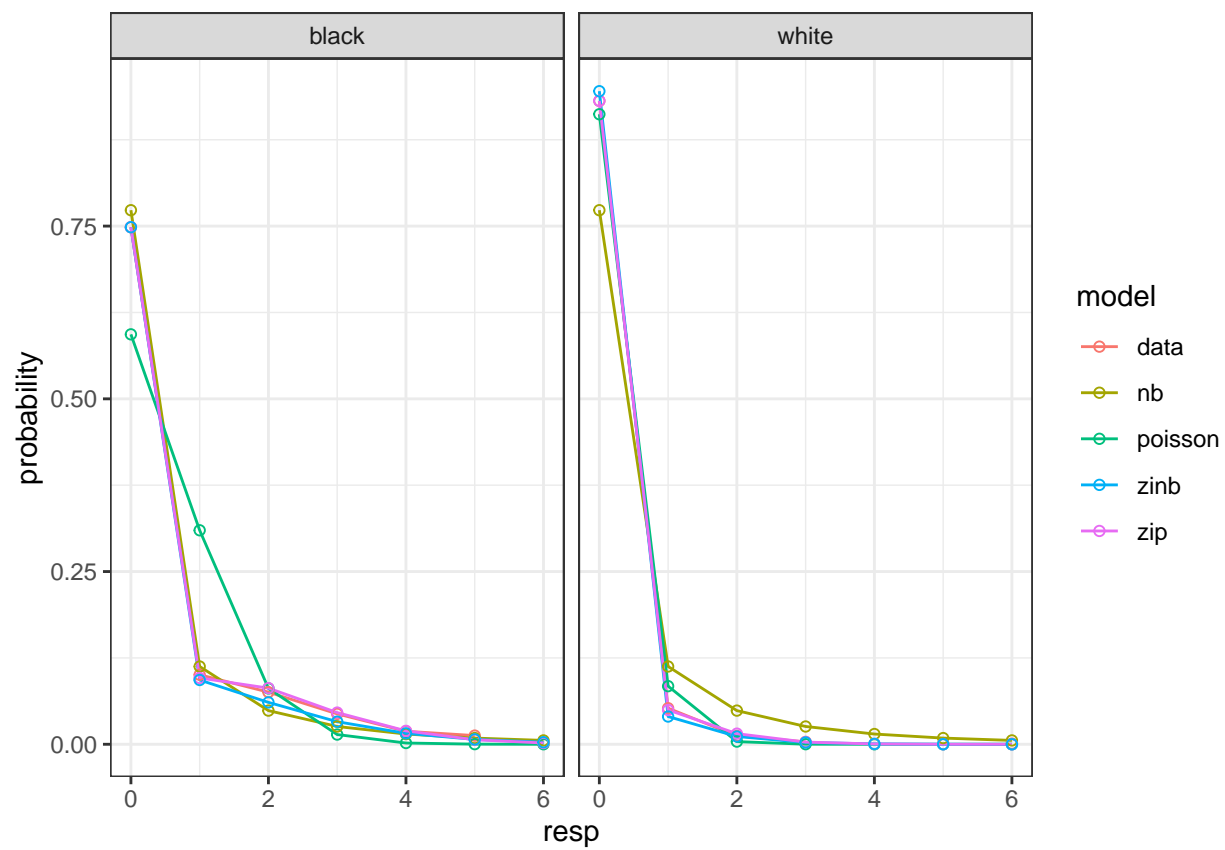
Figure 4: Graphical comparison of the sample probabilities and the predicted probabilities.