Generalized Linear Model

Group # 01

SEMINAR 2 Homicide Victims

Ricardo Castañeda r0731529

Qianli Fan r0775346

Butynets Mariia r0771332

Lieven Govaerts q0152493

Meng Wang r0767603

ZHANG Yanyi r0731121

Ruiman Zhong r0767577

Kendall Brown r0773111

PROBLEM DESCRIPTION

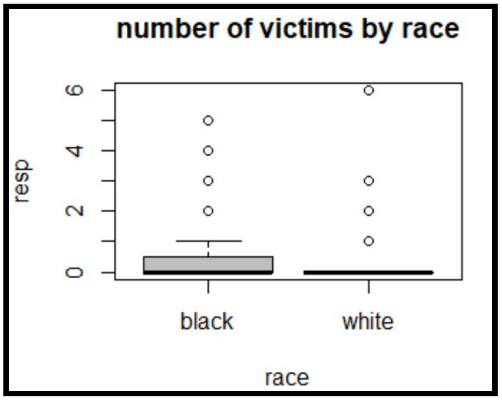
The data are from a survey of 1308 people in which they were asked how many homicide victims they know. The variables are:

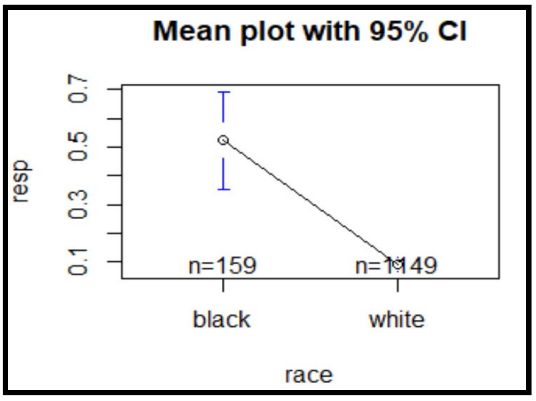
- ***resp**:* the number of victims the respondent knows (ranging from 0 to 6);
- ***race*:** the race of the respondent (159 black and 1149 white).

Scientific question: Does race help explain how many homicide victims a person knows?

1. EXPLORATION OF THE DATA

To have a better understanding of the data, descriptive statistics are extracted. Therefore, outliers could be tested and possible correlations could be examined. The first conclusion is that there is no missing data for both *race* and *resp*.





race	mean	var
black	0.52201258	1.1498288
white	0.09225413	0.1552448

A Box plot and a mean plot are created to see the effects of race on the number of homicide victims a respondent knows. According to the plots above, there seems to be a distinction between black and white people. Black people have a higher mean count than white people, and the variability of counts for the black population is larger than that for the white community. Meanwhile, the numerical analysis coincides with the graphical analysis. Note that for each race the sample variance roughly the double of the mean. It appears there is overdispersion that will be covered later.

2. POISSON MODEL

It is a common approach to start modeling counts with Poisson regression (here we count the number of victims a respondent knows, which are whole numbers greater than or equal to 0),

which assumes that the mean and variance of the response variable (denoted as λ) are the same. In our case, the response variable is the expected number of victims. We want to model the log-linear relationship between the race of the respondent (denoted as x) and the expected number of victims the respondent knows (denoted as y).

Theoretical model:

The systematic part of the (log-linear) Poisson regression model is given by

$$\ln[E(y|x)] = \beta 0 + \beta 1 * x$$

The distribution part is given by: y ~ Poisson(λ). After fitting the model to the data, we get the estimated model as: ln[E(y|x)] = -0.6501 - 1.7331*X, and the reference group for the race is the black respondent.

```
Call:
glm(formula = resp ~ race, family = poisson(link = "log"), data = victim)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0218  -0.4295  -0.4295  -0.4295   6.1874

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.6501     0.1098  -5.922 3.17e-09 ***
racewhite    -1.7331     0.1466 -11.825  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 962.80  on 1307  degrees of freedom
Residual deviance: 844.71  on 1306  degrees of freedom
AIC: 1122

Number of Fisher Scoring iterations: 6
```

According to the summary output of Poisson regression using the glm() function, at 5% significance level, the race has a significant effect on the expected number of victims the respondents know (p-value is almost zero). Moreover, it appears that white people are less likely to know someone who was a victim of homicide since the sign of the coefficient for the race is negative.

However, there seems to be some problems with the model as we can see that the residual deviance (844.71; which should be small for good fit of the model) is close to the Null deviance (962.80), which indicates that the ordinary Poisson model can’t capture the main variability of the dataset. Furthermore, there seems to be an overdispersion problem: as we discussed in the numerical analysis, the sample variance for each race is larger than its mean (Var(y) > E(y)). We will explore it later in detail with a formal overdispersion test and try to address this issue.

2.1 Risk ratio and corresponding confidence interval

We can calculate the risk ratio and its corresponding confidence interval based on the coefficients we estimated. The risk ratio for a race in which the reference group is black is 0.177, which means the average number of victims a white respondent knows is 0.177 times the average number of victims a black respondent knows; the average number of victims is less for white respondents than that for black respondents. The corresponding 95% confidence interval of the relative risk ratio is (0.133, 0.236), which means that we are 95 % confident that this range contains the risk ratio.

	RR	2.5 %	97.5 %
(Intercept)	0.522	0.418	0.643
racewhite	0.177	0.133	0.236

2.2 The ratio of the means of the response for each race

According to the model, the mean of the response for black respondent and white respondent can be calculated as follows:

E(y|x=0: black)=□^(-0.651)=0.522
E(y|x=1: white)=□^(-0.651-1.7331)=0.092

Therefore, the ratio of the means of the response for each race should be E(y|x=0: black)/E(y|x=1: white), which equals 5.659. It can be concluded that the average number of victims a black respondent knows is 5.659 times the average number of victims a white respondent knows.

2.3 Predictions of the models for each race

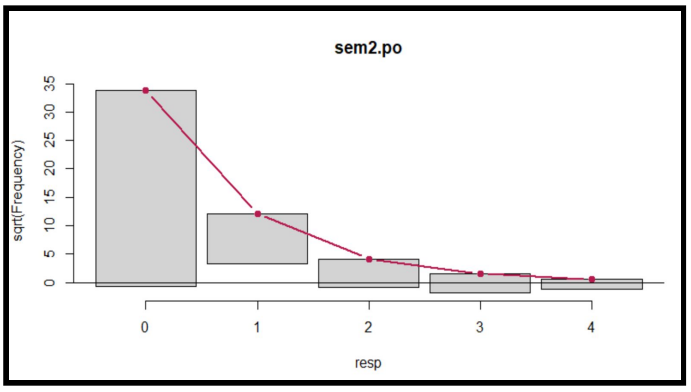
The number of victims a white/black person knows can be predicted by our model, which are the same as previous results as the mean of the response for each race. The numbers of victims we predict for white one and black one are 0.092 and 0.522 respectively. This states the count of known victims for white people is distributed as a Poisson distribution with mean and variance equal to 0.092, while the count of known victims for the black population is distributed as a Poisson distribution with mean and variance equal to 0.522.

2.4 Goodness of fit tests of the Poisson model

To apply the Poisson regression model, we should first do the goodness of fit test. Firstly, we did the Pearson Chi-Square test for this model: the Pearson residuals sum = 2279.87, and the p-value is 0, which means the model cannot capture the main features of the problem. While the deviance test shows that Residual deviance = 844.70 with degrees of freedom 1306, and the p-value is 1, which indicates that the model is suitable for us to predict the number of

victims a black/white knows. Since these values are not the same, we cannot conclude the goodness of fit of the model.

To decide whether our model is good to fit data or not, we need other evidence to help us out. We can use a rootogram to visualize the fit of a count regression model.



In the picture above, the red curved line is the theoretical Poisson fit. “Hanging” from each point on the red line is a bar that represents the difference between expected and observed counts. The counts have been transformed with a square root transformation to prevent smaller counts from getting obscured and overwhelmed by larger counts. In this case, a bar hanging below 0 indicates underfitting, while a bar hanging above 0 indicates overfitting. We see a great deal of underfitting for counts 0, 2, 3, 4 and higher and massive overfitting for the 1 count.

Apart from that, as we discussed above, the overdispersion seems to exist in the data because the variance of “resp” for each race is roughly two times the mean. The overdispersion test detected a larger variability in the data then we would expect given the Poisson model. The dispersion parameter is 1.75, which means that the variance of the response variable is 75 % larger than the mean. Based on what we discussed above (conflicting goodness-of-fit tests results and the presence of the overdispersion), we cannot conclude that the Poisson model is a suitable model for us to predict the results. So, we try to find another way to adjust our model.

```
        DHARMa nonparametric dispersion test via mean deviance residual fitted vs. simulated-refitted

data:  sim.model
dispersion = 1.7485, p-value < 2.2e-16
alternative hypothesis: two.sided
```

3. NEGATIVE BINOMIAL MODEL

One of the ways to take into account the overdispersion in our data is to build the Negative Binomial Model. Under this model, we assume that the response variable (numbers of victims) comes from the Negative Binomial distribution instead of the Poisson distribution. This allows us to consider the different relationship between the variance and the mean, namely that the variance of the negative binomial distribution is a quadratic function of its mean and it also involves an additional parameter (theta). This relationship can be expressed in the following way: variance of the response variable = E(y) + 1/theta * E(y)^2.

After we fitted the model, we can conclude that the estimates of the parameters are the same as for the Poisson model (-0.65 for the intercept, -1.73 for the race parameter). However, the standard errors are larger compared to the Poisson model, indicating that we have more uncertainty in our estimates (0.21 and 0.24). The race seems to have a significant effect on the expected number of victims.

```
Call:
glm.nb(formula = resp ~ race, data = victim, init.theta = 0.2023119205,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7184  -0.3899  -0.3899  -0.3899   3.5072

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.6501     0.2077  -3.130  0.00175 **
racewhite    -1.7331     0.2385  -7.268 3.66e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.2023) family taken to be 1)
```

3.1 Estimated variance versus observed variance

The model estimates the variance for the expected number of victims which are given in the table below. The dispersion parameter (theta) is 0.2023;

	White	Black
Observed variance	0.16	1.15
Estimated variance	0.13	1.87

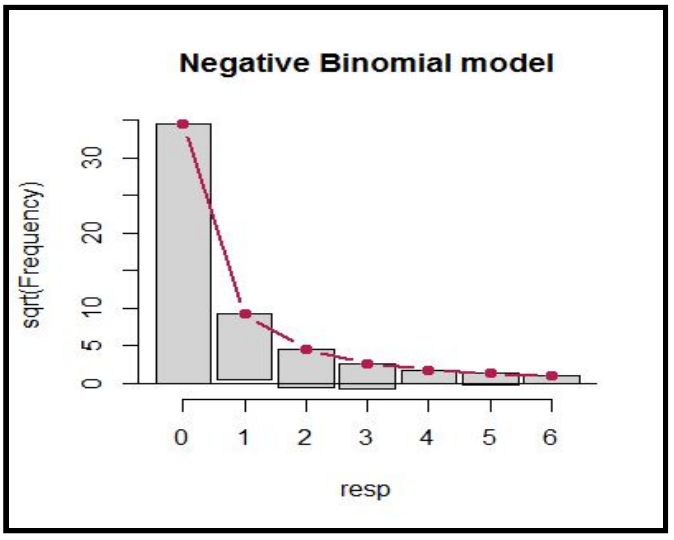
For white respondents, the estimated variance is slightly lower compared to the observed one; while for black respondents, the estimated variance is somewhat higher.

3.2 The goodness-of-fit tests

In the Pearson Chi-squared test, we reject the null hypothesis, which states that the current model is well specified (the expected number of victims can be calculated correctly based on the Negative Binomial distribution), since the sum of the Pearson residuals equals 1424.027 with p-value = 0.012. We also use the residual deviance to test whether the null hypothesis, which states that the current model provided an adequate fit for the data, should be rejected or not. Since the deviance is 412.59 with p-value 1, we can't reject the null hypothesis. Moreover, the residual deviance is less compared to the null deviance, indicating that adding covariate (race) to the model improves the fit of the model. The Pearson Chi-squared test and (Residual) Deviance test provide different results on the model adequateness of fit.

3.3 Visual exploration of fit

We use a rootogram to visualize the fit of the model. In the picture below, the red curved line is the theoretical Negative Binomial fit. We notice some improvement compared to the Poisson model, namely there is much less overfitting for 1 count, underfitting for 0 and 4 counts is diminished and underfitting for 2, 3 is reduced.



3.4 AIC

Now we want to compare the AIC values for the Poisson and Negative Binomial models. The AIC of the negative binomial model decreased to 1001.798, which shows that it has improved compared to the initial Poisson model we fitted.

Models	**df**	**AIC**
Poisson model	2	1121.990
Negative Binomial Model	3	1001.798

4. QUASI-LIKELIHOOD MODEL

Another type of model that can be fit in case of overdispersion is the Quasi-likelihood model. For this model, the mean and variance functions are specified separately. And it can be assumed that there is a linear relationship between the mean and the variance of the response variable, but we do not expect them to be the same. And we do not specify the type of distribution for the response variable.

The systematic part of the model is given by: $log(\mu) = \beta_0 + \beta_1 x_1 + \epsilon_i$, $var(y) = \omega\mu$ (the variance of the response variable = dispersion parameter (theta) * expected value of the response variable).

The estimates for the parameters are the same as for the Poisson and the Negative Binomial models. The standard errors for the estimates are larger compared to the Poisson, but smaller compared to the Negative Binomial model. The race seems to have a significant effect on the expected number of victims. And the variance is about 75% larger than the mean.

```
Call:
glm(formula = resp ~ race, family = quasipoisson, data = victim)

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)
```

5. DISCUSS ALL THE RESULTS

5.1 Comparing the estimates of the parameters and standard errors

	Po	QL	NB	se.Po	se.QL	se.NB
(Intercept)	-0.6501	-0.6501	-0.6501	0.1098	0.1450	0.2077
racewhite	-1.7331	-1.7331	-1.7331	0.1466	0.1937	0.2385

Comparing all three models (Poisson, Quasi-likelihood and Negative Binomial), as shown above in the graph, we may conclude that the estimates for the parameters are the same for all models and standard errors of the parameter estimates are larger for Quasi-likelihood and Negative Binomial models compared to Poisson model. The larger standard errors indicate that we have more uncertainty in our estimate. In other words, the Quasi-likelihood and Negative Binomial models take into account the variability that exists in the data.

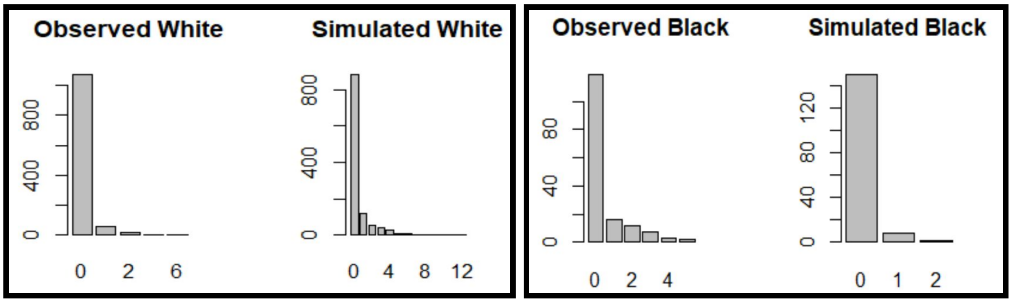
5.2 GOF tests

	Pearson Chi-square Test(P value)	Deviance Test(P-value)
Poisson model	0	1
Negative Binomial Model	0.01202883	1

As the table shows, the results of the goodness-of-fit tests are the same for Poisson and Negative Binomial models. The Pearson Chi-squared tests provide the evidence that both models do not capture the main features of the problem we study, and the deviance tests do not provide the evidence that the models do not fit the data well (the models almost get the same p values for Pearson Chi-square test and deviance test). The obvious difference is that the negative binomial model has a higher p-value of 0.012 than the other two, which means that it has less significant power to reject H0 (our model is well specified).

5.3 Further exploration of the Negative Binomial model

Based on all the analyses and comparisons we have done, it seems that the negative binomial model works better to adjust the overdispersion problem. Besides, the variances given by the negative binomial model are also close to the observed variances (see part 3).
To gain further insights into our negative binomial model, let's use its parameters to simulate data and compare the simulated data to the observed data. We simulate the same number of observations as we have in our original data.



As we can see from the plots, the simulated data are similar to the observed data, again giving us confidence in choosing negative binomial regression to model this data. These plots also demonstrate the conditional nature of our model and answer our scientific question. That is, **the negative binomial distribution of the counts depends on race, and in other words, the race variable helps explain how many homicide victims a person knows.**