Poisson Regression

2019-04-14

In this set of notes, you will learn how Poisson regression models can be used to model outcome variables that are counts.

Dataset and Research Question

In this set of notes, we will use data from the file *stop-and-frisk.csv* (see the data codebook here). These data include the number of "stop and frisk" stops made by New York City police officers by ethnic group. It also includes some other-precinct-level attributes.

```
# Load libraries
library(AICcmodavg)
library(broom)
library(corrr)
library(dplyr)
library(ggplot2)
library(readr)
library(sm)
library(tidyr)

# Read in data
frisk = read_csv(file = "~/Documents/github/epsy-8252/data/stop-and-frisk.csv")

# View data
head(frisk)
```

```
# A tibble: 6 x 6
  stops population past_arrests precinct ethnicity crime
  <dbl>      <dbl>        <dbl>    <dbl> <chr>     <chr>
1    75       1720          191        1 Black     Violent
2    37       1368           62        1 Hispanic  Violent
3    26      23854          135        1 White     Violent
4    36       1720           57        1 Black     Weapons
5    39       1368           27        1 Hispanic  Weapons
6    32      23854           16        1 White     Weapons
```

We will use these data to explore whether or not there is evidence that minorities were detained with "stop and frisk" more frequently than whites.

Poisson Distribution

The Poisson distribution is a probability distribution that is often used to model the variation in count data. (Count data takes on integer values greater than or equal to zero; $y_i \in \{0, 1, 2, 3, \ldots\}$.) Specifically, the Poisson distribution "expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant rate and independently of the time since the last event" (Wikipedia).

In the Poisson model, each case i corrsponds to a setting (typically a time interval or spatial location) in which y_i events are observed. For example, consider a researcher who wants to model the number of accidents in Dinkytown. That person collects data on the number of accidents (y) occurring at on each street (i) over the course of a given year (fixed interval of time). Other example uses of the Poisson distribution could be to model:

- The number of patients arriving at HCMC between 10:00 p.m. and 12:00 a.m.;
- The number of mood changes observed in children with Rett Syndrome over the course of an hour;
- The number of spelling errors made on a test of 25 words.

The probability of observing k events in an interval is given by the equation:

$$P(k \text{ events in the interval}) = e^{-\lambda} \times \frac{\lambda^k}{k!}$$

where λ is the average number of events in the given interval.

As an example, suppose you gave students in a class a weekly spelling test of 25 items, and you knew that these students averaged four spelling errors per test. What is the probability that a student would have seven spelling errors on her test? We could model this using a Poisson distribution using the following values:

- $\lambda = 4$
- $k = 7$

$$P(k = 7) = e^{-4} \times \frac{4^7}{7!}$$

```
exp(-4) * (4^7) / factorial(7)
```

```
[1] 0.0595
```

The probability of seeing a student make 7 errors, given the average rate of errors is 2 per test, is 0.06; a fairly unlikely event. We can also use the `dpois()` function to compute that same probability.

```
dpois(x = 7, lambda = 4)
```

```
[1] 0.0595
```

The λ parameter completely defines the shape, mean, and variance of the Poisson distribution. The figure below shows how the probability function differs based on the mean (or rate) parameter.

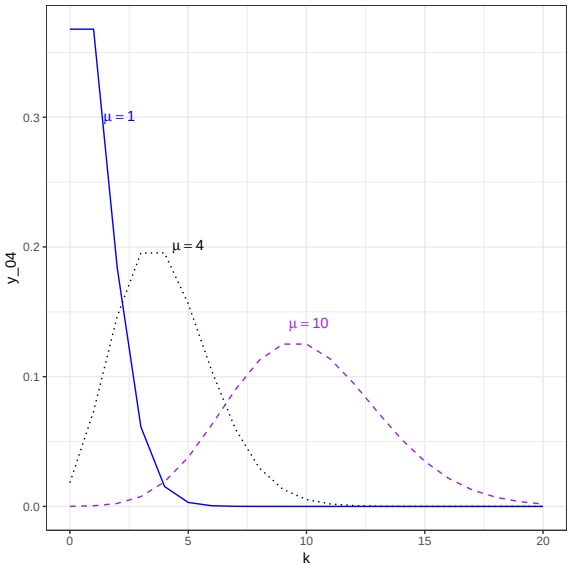


Figure 1: The probability mass function for Poisson distributions with means of 1 (blue solid line), 4 (black dotted line), and 10 (purple dashed line).

In the Poisson distribution both the mean and variance are the same. Thus in the Poisson distribution with $\lambda = 4$, both the mean and variance are equal to 4. One key assumption of the Poisson distribution is that the rate of occurrence is constant over the interval. The rate cannot be higher in some intervals and lower in other intervals. Another assumption is that the occurrence of any one event does not affect the probability that any other event will occur (independence). For example, data collected on the number of students who arrive at the Coffman bus stop per minute will likely not follow a Poisson distribution, because the rate is not constant; buses tend to run only at certain times of the day and have higher rates at peak hours. Furthermore the arrivals of individual students are not independent (students tend to arrive in groups).

Poisson Regression Using a Generalized Linear Model

We can use a generalized linear model to predict variation in an outcome that consists of count data. Remember, these models consist of three components:

- A **linear function describing the structure between the predictors,** $X_1, X_2, \ldots, X_k$. This structure can be additive (main-effects) or multiplicative (interactions) in nature, and can include transformations of the predictors, polynomial terms, dummy coded predictors, etc.
- A **link function** which transforms the mean of the outcome variable to the specified linear set of predictors. This function needs to be mathematically smooth (no gaps or jumps) and invertible (we can backtransform). For example if our link function is $g(\cdot)$, then
- A **random component** specifying the conditional distribution of the response variable, Y_i, given the predictors in the model. This distribution is either a member of the *exponential family* of distributions or from the *multivariate exponential family* of distributions.

To fit a generalized linear model to count data it is common to use the log transformation to link the mean of the outcome and the set of linear predictors:

$$\ln(\mu_i) = \beta_0 + \beta_1(X_{i1}) + \beta_2(X_{i2}) + \ldots + \beta_k(X_{ik})$$

In addition, we typically specify the random errors as a Poisson distribution.

$$\epsilon_i \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$$

The syntax to fit the Poisson regression model using `glm()` is

$$\texttt{glm}(\text{y} \sim 1 + \text{x}, \texttt{data} = \text{dataframe}, \texttt{family} = \texttt{poisson}(\texttt{link} = \texttt{log})$$

Modeling the "Stop and Frisk" Data

To begin the analysis, we will fit a model that only includes the focal predictor of ethnicity. Because the primary research question is focused on whether minority groups are being disadvantaged we will set the reference group to `White` rather than letting R pick the reference group alphabetically. To do this we use the `relevel()` function. This function only works on factors, so we also need to coerce `ethnicity` into a factor.

```
glm.1 = glm(stops ~ 1 + relevel(factor(ethnicity), ref = "White"),
            data = frisk, family = poisson(link = "log"))
summary(glm.1)
```

```
Call:
glm(formula = stops ~ 1 + relevel(factor(ethnicity), ref = "White"),
    family = poisson(link = "log"), data = frisk)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-21.58  -10.12   -4.95    1.82   63.61

Coefficients:
                                                Estimate Std. Error
(Intercept)                                      4.03566    0.00768
relevel(factor(ethnicity), ref = "White")Black   1.41428    0.00856
relevel(factor(ethnicity), ref = "White")Hispanic  0.96657    0.00902
                                                z value
(Intercept)                                         526
relevel(factor(ethnicity), ref = "White")Black      165
relevel(factor(ethnicity), ref = "White")Hispanic   107
                                                            Pr(>|z|)
(Intercept)                                     <0.0000000000000002 ***
relevel(factor(ethnicity), ref = "White")Black  <0.0000000000000002 ***
relevel(factor(ethnicity), ref = "White")Hispanic <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 182217  on 899  degrees of freedom
Residual deviance: 147656  on 897  degrees of freedom
AIC: 153000

Number of Fisher Scoring iterations: 5
```

The fitted equation is:

$$\ln(\hat{\mu}_i) = 4.04 + 1.41(\text{Black}_i) + 0.97(\text{Hispanic}_i)$$

Similar to models in which we log-transformed the outcome, we can gain better interpretations by exponentiating (back-transforming) the coefficients from the Poisson model.

$$\begin{aligned} e^{\ln(\hat{\mu}_i)} &= e^{4.04+1.41(\text{Black}_i)+0.97(\text{Hispanic}_i)} \\ \hat{\mu}_i &= 56.58 \times 4.11(\text{Black}_i) \times 2.63\text{Hispanic}_i) \end{aligned}$$

Using R, we can quickly back-transform the coefficients.

```
exp(coef(glm.1))
```

```
                                     (Intercept)
                                           56.58
   relevel(factor(ethnicity), ref = "White")Black
                                            4.11
relevel(factor(ethnicity), ref = "White")Hispanic
                                            2.63
```

- The predicted average number of "stop and frisk" stops for Whites (the reference group) across the 75 precincts is 56.58.
- Blacks are stopped, on average, 4.11 times as often as Whites with "stop and frisk" ($p < .001$).
- Hispanics are stopped, on average, 2.63 times as often as Whites with "stop and frisk" ($p < .001$).

Rates of Occurence: Including an Offset in the Model

In most Poisson analyses, the counts are interpreted relative to some baseline, or in the Poisson regression parlance, *exposure*. (Exposure is a term that initially comes from the health fields, but is used in Poisson modeling.) For example, the number of "stop and frisk" stops that occur in a precinct for each minority group. Or the number of accidents that occur on each street in Dinkytown.

In some analyses it may be more beneficial to analyze the *rate of occurrence* rather than the number of cases. To do this we let y_i be the number of cases observed out of a_i cases that are exposed to the risk. For example, it may make sense in our analysis of the "stop and frisk" data to analyze the rate of occurrence relative to the number of arrests of the different ethnic groups made in the previous year.

To do this, we include the $\ln(\text{exposure})$ as a term in the model. This changes our link function to:

$$\ln(\mu_i) = \beta_0 + \beta_1(X_{i1}) + \beta_2(X_{i2}) + \ldots + \beta_k(X_{ik}) + \ln(a_i)$$

This last term ($\ln(a_i)$) is referred to as the *offset*, and the $\ln(\mu_i)$ now represents the logarithm of the mean rate of occurrence (rather than the logarithm of the mean occurrence).

To include an offset in the `glm()` function we add the argument `offset=` and provide this with a log-transformed variable that includes the number of cases (counts) exposed to the risk; in our example, the logarithm of the number of people from each minority group arrested in each precinct the previous year. (Note that we need to add 1 before taking the log because we have values of 0 in the `past_arrests` variable.)

```
glm.1.2 = glm(stops ~ 1 + relevel(factor(ethnicity), ref = "White"), data = frisk,
              family = poisson(link = "log"), offset = log(past_arrests+1))
summary(glm.1.2)
```

```
Call:
glm(formula = stops ~ 1 + relevel(factor(ethnicity), ref = "White"),
    family = poisson(link = "log"), data = frisk, offset = log(past_arrests +
        1))

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-47.84   -5.96    0.43    7.17   64.21

Coefficients:
                                                Estimate Std. Error
(Intercept)                                     -0.75798    0.00768
relevel(factor(ethnicity), ref = "White")Black   0.16751    0.00856
relevel(factor(ethnicity), ref = "White")Hispanic  0.23611    0.00902
                                                z value
(Intercept)                                       -98.8
relevel(factor(ethnicity), ref = "White")Black     19.6
relevel(factor(ethnicity), ref = "White")Hispanic    26.2
                                                        Pr(>|z|)
(Intercept)                                     <0.0000000000000002 ***
relevel(factor(ethnicity), ref = "White")Black  <0.0000000000000002 ***
relevel(factor(ethnicity), ref = "White")Hispanic <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 183267  on 899  degrees of freedom
Residual deviance: 182558  on 897  degrees of freedom
AIC: 187902

Number of Fisher Scoring iterations: 6
```

Back-transforming the coefficients:

```
exp(coef(glm.1.2))
```

```
                                     (Intercept)
                                           0.469
   relevel(factor(ethnicity), ref = "White")Black
                                           1.182
relevel(factor(ethnicity), ref = "White")Hispanic
                                           1.266
```

We now interpret these values relative to the number of arrests in the previous year. Values above 1 indicate that there are more arrests for that ethnicity group relative to the previous year and values below 1 indicate there are fewer arrests relative to the number the previous year. For example,

- The predicted average number of "stop and frisk" stops for Whites (the reference group) across the 75 precincts is 0.46 times the proportion of arrests for whites in the previous year.
- Blacks are stopped, on average, 1.18 times as often as Whites with "stop and frisk" relative to the arrest rate from the previous year ($p < .001$).
- Hispanics are stopped, on average, 1.27 times as often as Whites with "stop and frisk" relative to the arrest rate from the previous year ($p < .001$).

Interpret as Percentages Directly (No Back-Transforming)

We could have answered our RQ without back-transforming. The fitted equation was:

$$\ln(\hat{\mu}_i) = -0.76 + 0.17(\text{Black}_i) + 0.24(\text{Hispanic}_i)$$

Since the Poisson model uses the natural logarithm in the link function, we can take advantage of that and interpret coefficients directly as percentage differences. For example,

- Blacks are stopped, on average, 17% more often than Whites with "stop and frisk" relative to the arrest rate from the previous year ($p < .001$).
- Hispanics are stopped, on average, 24% more often than Whites with "stop and frisk" relative to the arrest rate from the previous year ($p < .001$).

Technically, we should use the expression

$$e^{\hat{\beta}-1}$$

to obtain more accurate estimates. For example,

$$\begin{aligned} \textbf{Black}: \ & e^{0.17} - 1 = 0.185 \\ \textbf{Hispanic}: \ & e^{0.24} - 1 = 0.27.1 \end{aligned}$$

But, the direct interpretations are close enough approximations in most cases. (The bigger the coefficients are, the further the accurate estimates are from the direct coefficient values.)

Including Covariates

Before we conclude that there are a disproportionate number of "stop and frisk" stops for minorities, we may want to make a stronger argument by controlling for other relevant factors. For example, one argument may be that these differences are really due to precinct-level differences. Let's see if what happens when we control for precinct. To do this, we include precinct as a factor (after all, it is a categorical variable).

```
glm.2 = glm(stops ~ 1 + relevel(factor(ethnicity), ref = "White") + factor(precinct),
            data = frisk, family = poisson(link = "log"), offset = log(past_arrests+1))

summary(glm.2)
```

```
Call:
glm(formula = stops ~ 1 + relevel(factor(ethnicity), ref = "White") +
    factor(precinct), family = poisson(link = "log"), data = frisk,
    offset = log(past_arrests + 1))

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-39.42   -6.05   -1.15    5.18   58.39

Coefficients:
                                                Estimate Std. Error z value             Pr(>|z|)
(Intercept)                                    -1.80706    0.05152  -35.08 < 0.0000000000000002 ***
relevel(factor(ethnicity), ref = "White")Black  0.42213    0.00942   44.81 < 0.0000000000000002 ***
relevel(factor(ethnicity), ref = "White")Hispanic  0.43035    0.00958   44.94 < 0.0000000000000002 ***
factor(precinct)2                              -0.14840    0.07403   -2.00              0.04501 *
factor(precinct)3                               0.56413    0.05676    9.94 < 0.0000000000000002 ***
      :                                             :          :       :                    :
factor(precinct)73                              0.99599    0.05359   18.59 < 0.0000000000000002 ***
factor(precinct)74                              1.15331    0.05802   19.88 < 0.0000000000000002 ***
factor(precinct)75                              1.54060    0.07572   20.35 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 183267  on 899  degrees of freedom
Residual deviance: 140788  on 823  degrees of freedom
AIC: 146281

Number of Fisher Scoring iterations: 6
```

Since the estimates are large, it probably behooves us to use the more accurate estimates:

$$\textbf{Black}: \ e^{0.42} - 1 = 0.525$$
$$\textbf{Hispanic}: \ e^{0.430} - 1 = 0.537$$

After controlling for precinct-level differences, the effect is even larger!

- Blacks are stopped, on average, 52.5% more often than Whites with "stop and frisk" relative to the arrest rate from the previous year, after controlling for precinct differences ($p < .001$).
- Hispanics are stopped, on average, 53.7% more often than Whites with "stop and frisk" relative to the arrest rate from the previous year, after controlling for precinct differences ($p < .001$).

Assumption Checking

Checking the assumptions, we have to be careful. This is because we don't expect to see the same patterns in the residual plots. For example, when examining the density plot of the residuals, we now expect that the residuals will be Poisson distributed (not normally distributed). Also when examining a scatterplot of the residuals versus the fitted values, we expect that the variance will not be constant; remember the variance of Posisson distributed variable should be equal to its mean. This implies that the variance should increase for higher fitted values (fan-shaped).

In examining assumptions for generalized models, it is common to transform the residuals. There are many ways to do this, but one common method for assumption checking is to compute the *Pearson residual*. The Pearson residual essentially divides the residual value by its conditional standard deviation (standardizing it). The advantage is that the conditional distributions of residuals are now all on the same scale and can be evaluated for homoscedasticity.

To obtain the Pearson residuals, we include the argument `type.residual = "pearson"` in the `augment()` function. The column labelled `.resid` now includes the Pearson residuals rather than the raw residuals.

```
# Get Pearson residuals
out_2 = augment(glm.2, type.residual = "pearson")
head(out_2)
```

```
# A tibble: 6 x 11
  stops relevel.factor.~ factor.precinct. X.offset. .fitted .se.fit .resid
  <dbl> <fct>            <fct>                <dbl>   <dbl>   <dbl>  <dbl>
1    75 Black            1                     5.26    3.87  0.0510  3.88
2    37 Hispanic         1                     4.14    2.77  0.0512  5.29
3    26 White            1                     4.91    3.11  0.0515  0.778
4    36 Black            1                     4.06    2.68  0.0510  5.64
5    39 Hispanic         1                     3.33    1.96  0.0512 12.0
6    32 White            1                     2.83    1.03  0.0515 17.5
# ... with 4 more variables: .hat <dbl>, .sigma <dbl>, .cooksd <dbl>,
#   .std.resid <dbl>
```

```
# Residual plot
ggplot(data = out_2, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  geom_smooth() +
  theme_bw() +
  xlab("Fitted values") +
  ylab("Pearson residuals")
```

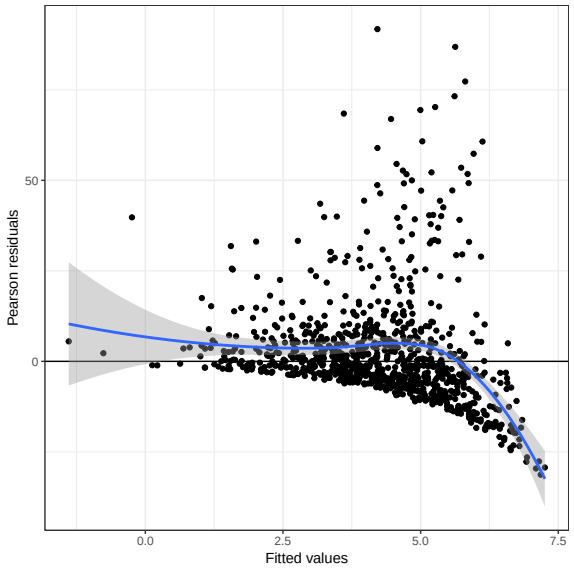


Figure 2: Plot of the Pearson residuals versus the fitted values for the Poisson regression model that inlcudes ethnicity and precinct as predictors of the number of "stop and frisk" stops. The number of past arrests was used as an offset in the model.

We can also use the `residualPlot()` function from the **car** package to directly create this plot.

```
library(car)
residualPlot(glm.2)
```

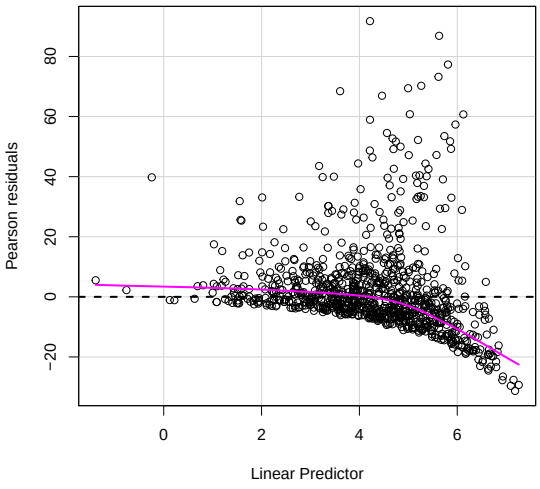


Figure 3: Plot of the Pearson residuals versus the fitted values using the 'residualPlot()' function from the **car** package.

The residual plot indicates that the model's assumptions are likely violated. The residuals do not have constant variance after scaling them by their conditional standard deviations. This likely indicates that the data are not Poisson distributed, and thus need to be modeled using a different distribution.

Overdispersion

Looking at the original data (conditioned on ethnicity and precinct) we may have a clue as to the problem. Namely that the conditional mean and variances are not equal as we would expect in a Poisson distribution. Here those conditional values are shown for the Precincts 1, 2, and 3 (although in practice you would looks at all of them).

```
frisk %>%
  filter(precinct %in% c(1:3)) %>%
  group_by(ethnicity, precinct) %>%
  summarize(
    M = mean(stops),
    V = var(stops)
  )
```

```
# A tibble: 9 x 4
# Groups:   ethnicity [3]
  ethnicity precinct     M      V
  <chr>        <dbl> <dbl>  <dbl>
1 Black            1  50.5  828.
2 Black            2  33    864
3 Black            3 188   9149.
4 Hispanic         1  25.5  276.
5 Hispanic         2  36    907.
6 Hispanic         3 110.  1462.
7 White            1  20.2  110.
8 White            2  17.8   80.2
9 White            3 102.   164.
```

The variances are much larger than the means. This is referred to as *overdispersion*. If the variances were lower than the means, the data would be underdispersed. We can also look for evidence of overdispersion by examining the ratio of the residual deviance and the residual degrees of freedom (from the `summary()` output). In general, if the ratio of these two values is greater than 1, there is likely overdispersion. Values less than 1 indicate underdispersion. In our output:

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 183267  on 899  degrees of freedom
Residual deviance: 140788  on 823  degrees of freedom
AIC: 146281

Number of Fisher Scoring iterations: 6
```

The residual deviance was 140,788 and the residual degrees of freedom was 823. This ratio is:

$$\frac{140788}{823} = 171$$

This is evidnce of sever overdispersion in the data. The are several methods for adjusting for overdispersion. Which you choose will ultimately be based on the residual fit; choose the method that has the best behaved residuals.

Quasilikelihood Methods

One common method of adjusting for overdispersion estimates the model using quasilikelihood. To fit the Poisson model using quasilikelihood, we change the `family=` argument in the `glm()` function from `family=poisson(link="log")` to `family=quasipoisson(link = "log")`.

```
# Fit the model using quasilikelihood estimates
glm.3 = glm(stops ~ 1 + relevel(factor(ethnicity), ref = "White") + factor(precinct), data = frisk,
            family = quasipoisson(link = "log"), offset = log(past_arrests+1))

summary(glm.3)
```

```
Call:
glm(formula = stops ~ 1 + relevel(factor(ethnicity), ref = "White") +
    factor(precinct), family = quasipoisson(link = "log"), data = frisk,
    offset = log(past_arrests + 1))

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-39.42   -6.05   -1.15    5.18   58.39

Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                      -1.8071     0.8288   -2.18   0.0295 *
relevel(factor(ethnicity), ref = "White")Black    0.4221     0.1516    2.79   0.0055 **
relevel(factor(ethnicity), ref = "White")Hispanic 0.4304     0.1541    2.79   0.0053 **
factor(precinct)2                                -0.1484     1.1910   -0.12   0.9009
factor(precinct)3                                 0.5641     0.9131    0.62   0.5369
      :                                              :          :       :       :
factor(precinct)73                                0.9960     0.8621    1.16   0.2483
factor(precinct)74                                1.1533     0.9334    1.24   0.2170
factor(precinct)75                                1.5406     1.2181    1.26   0.2063
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 259)

    Null deviance: 183267  on 899  degrees of freedom
Residual deviance: 140788  on 823  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
```

The coefficients from the fitted model can be interpreted in the same manner as those from the Poisson model.

$$\textbf{Black}: \; e^{0.4221} - 1 = 0.525$$
$$\textbf{Hispanic}: \; e^{0.4304} - 1 = 0.538$$

- Blacks are stopped, on average, 52.5% more often than Whites with “stop and frisk” relative to the arrest rate from the previous year, after controlling for precinct differences ($p < .001$).
- Hispanics are stopped, on average, 53.8% more often than Whites with “stop and frisk” relative to the arrest rate from the previous year, after controlling for precinct differences ($p < .001$).

Examining the Pearson residuals we find that the residuals from the quasi-Poisson model still are not well-behaved.

```
# Check the residuals
residualPlot(glm.3)
```

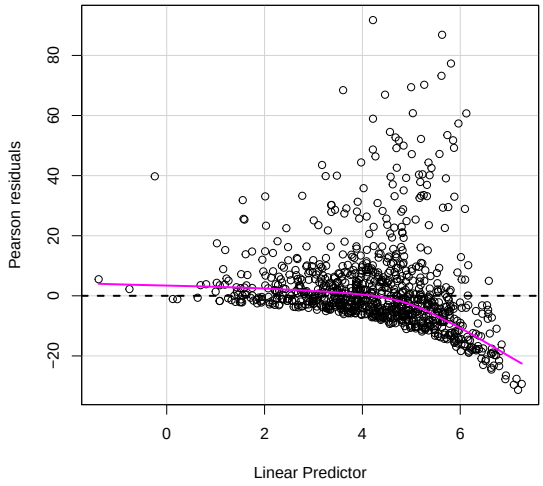


Figure 4: Plot of the Pearson residuals versus the fitted values for the quasilikelihood estimated Poisson regression model.

Negative Binomial Model

Another common method is to use a negative binomial model to fit the data rather than a Poisson model. This model is interpreted in the same way as the Poisson model, but allows an additional paramter to account for the overdispersion. To fit a negative binomial model using R, we need to use the `glm.nb()` function from the **MASS** package. Because this function always fits a negative binomial, we do not need to specify the family in this function.

```
library(MASS)
glm.4 = glm.nb(stops ~ 1 + relevel(factor(ethnicity), ref = "White") + factor(precinct) + offset(log(past_arrests+
               data = frisk)

summary(glm.4)
```

```
Call:
glm.nb(formula = stops ~ 1 + relevel(factor(ethnicity), ref = "White") +
    factor(precinct) + offset(log(past_arrests + 1)), data = frisk,
    init.theta = 1.082665365, link = log)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.855  -1.111  -0.413   0.342   3.603

Coefficients:
                                                Estimate Std. Error z value  Pr(>|z|)
(Intercept)                                    -0.997617   0.286326   -3.48   0.00049 ***
```

```
relevel(factor(ethnicity), ref = "White")Black     0.367423   0.079731    4.61 0.0000041 ***
relevel(factor(ethnicity), ref = "White")Hispanic  0.370574   0.079813    4.64 0.0000034 ***
factor(precinct)2                                 -0.529965   0.400237   -1.32   0.18546
factor(precinct)3                                  0.831212   0.396769    2.09   0.03618 *
      :                                               :          :         :         :
factor(precinct)73                                 0.344121   0.396461    0.87   0.38541
factor(precinct)74                                 0.591696   0.397381    1.49   0.13649
factor(precinct)75                                 0.451555   0.409594    1.10   0.27027
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1.08) family taken to be 1)

    Null deviance: 1286.5  on 899  degrees of freedom
Residual deviance: 1036.3  on 823  degrees of freedom
AIC: 10336

Number of Fisher Scoring iterations: 1


              Theta:  1.0827
          Std. Err.:  0.0475

 2 x log-likelihood:  -10180.2170
```

The coefficients from the fitted model can be interpreted in the same manner as those from the Poisson model.

$$\textbf{Black}: \ e^{0.367} - 1 = 0.443$$
$$\textbf{Hispanic}: \ e^{0.371} - 1 = 0.449$$

- Blacks are stopped, on average, 44.3% more often than Whites with "stop and frisk" relative to the arrest rate from the previous year, after controlling for precinct differences ($p < .001$).
- Hispanics are stopped, on average, 44.9% more often than Whites with "stop and frisk" relative to the arrest rate from the previous year, after controlling for precinct differences ($p < .001$).

Based on the Pearson residuals, the negative binomial indicates much better fit to the data than either the Poisson or quasi-Poisson models.

```
residualPlot(glm.4)
```

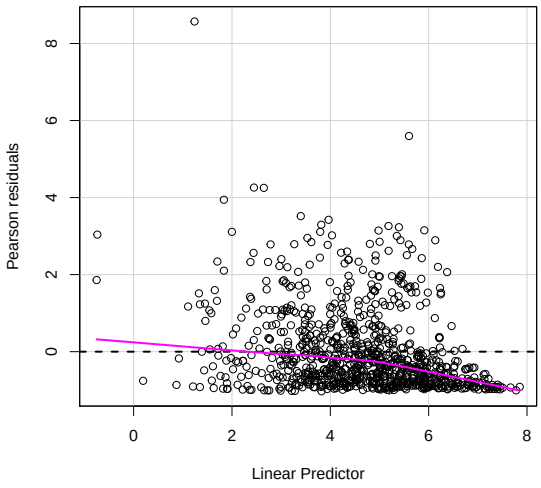


Figure 5: Plot of the Pearson residuals versus the fitted values for the negative binomial regression model.

Zero Inflated Models

One common issue in modeling count data is that there are too many zero values. For example, below we examine the number of stops by ethnicity.

```
ggplot(data = frisk, aes(x = stops)) +
  geom_histogram(color = "black", fill = "skyblue") +
  facet_wrap(~ethnicity) +
  theme_bw() +
  xlab('"Stop and frisk" stops') +
  ylab("Counts")
```

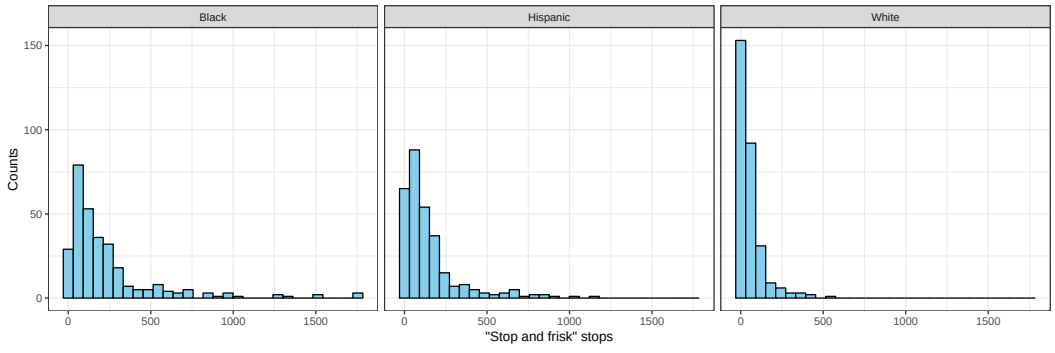


Figure 6: Histogram showing the number of "stop and frisk" stops by ethnicity.

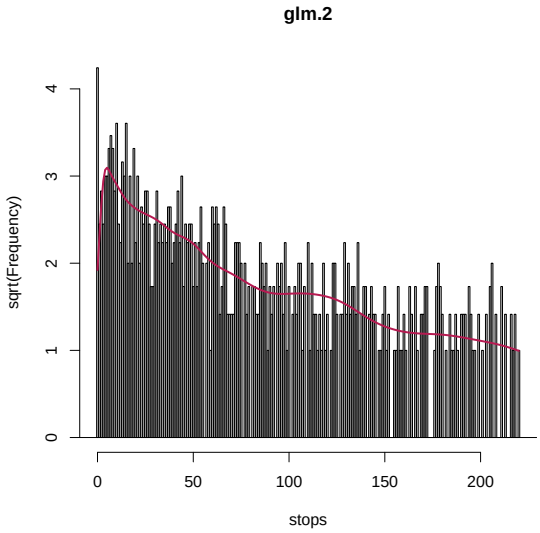
There seems to be more values of zero than would be expected in a Poisson distribution, especially for Whites. We can also examine this via a *standing rootogram*. This plot compares the fitted values of the Poisson or negative binomial model to the observed counts. We first need to install the **countreg** package from R-Forge using the following syntax.

```
# Install the countreg package
install.packages("countreg", repos="http://R-Forge.R-project.org")
```

Then we can use the `rootogram()` function with the argument `style="standing"` to plot a standing rootogram.

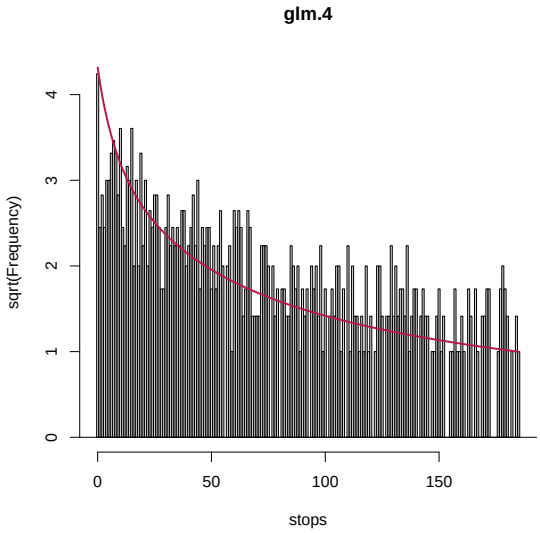
```
# Load the package
library(countreg)

# Standing rootogram of the fitted Poisson model
rootogram(glm.2, style = "standing")
```



The rootogram shows poor model fit in general, but especially at zero stops. The observed data (bars) is much higher than the expected counts given by the model (red line) at zero. What about the negative binomial model?

```
# Standing rootogram of the fitted negative binomial model
rootogram(glm.4, style = "standing")
```



In this model, the model seems to do a better job in general than the Poisson model; the model predicted values are much closer to the observed counts. Allowing for overdispersion also seemed to better model the number of zeros. Thus it does not seem necessary to use a zero-inflated model.

Zero-Inflated Models

Zero-inflated models essentially model an additional data generating process that is based on the probability of whether the count is a zero or not. Then the counts that are not zeros are modeled by a Poisson process. To fit a zero-inflated model use the `zeroinfl()` function from the **pscl** package. You will need to also specify the `distr=` argument using the family to use `poisson` or `negbin`). Here we fit a zero-inflated negative binomial model.

```
library(pscl)
glm.5 = zeroinfl(stops ~ 1 + relevel(factor(ethnicity), ref = "White") + factor(precinct) + offset(log(past_arrests
               data = frisk, dist = "negbin")

# Standing rootogram
rootogram(glm.5, style = "standing")
```

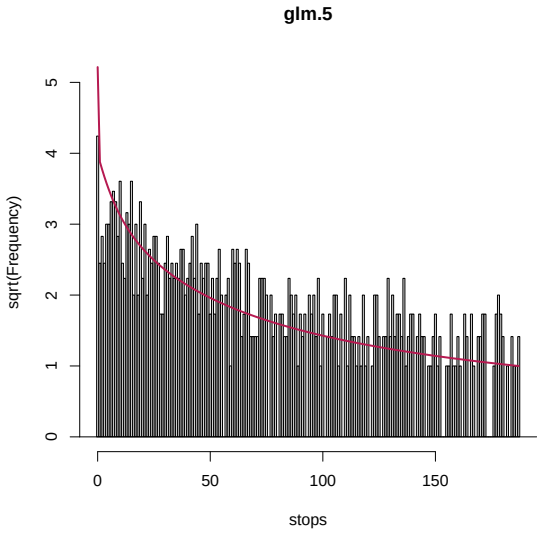


Figure 7: Plot of the Pearson residuals versus the fitted values for the negative binomial regression model.

As expected, the model predicted values at 0 are much more inflated. Here they are too inflated producing residual misfit.

Understanding the `summary()` Output

To help you understand the output from fitting the zero-inflated model, I will fit a simpler model (for pedagogical ease) that only includes the ethnicity predictor. I will also use the Poisson distribution to model the counts.

```
glm.example = zeroinfl(stops ~ 1 + relevel(factor(ethnicity), ref = "White") + offset(log(past_arrests+1)),
               data = frisk, dist = "poisson")
summary(glm.example)
```

```
Call:
zeroinfl(formula = stops ~ 1 + relevel(factor(ethnicity), ref = "White") +
    offset(log(past_arrests + 1)), data = frisk, dist = "poisson")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-35.257  -3.292   0.389   7.870 101.955

Count model coefficients (poisson with log link):
                                                Estimate Std. Error
(Intercept)                                     -0.74444    0.00768
relevel(factor(ethnicity), ref = "White")Black   0.15397    0.00856
relevel(factor(ethnicity), ref = "White")Hispanic  0.22382    0.00902
                                                z value
(Intercept)                                       -96.9
relevel(factor(ethnicity), ref = "White")Black     18.0
relevel(factor(ethnicity), ref = "White")Hispanic  24.8
                                                         Pr(>|z|)
(Intercept)                                     <0.0000000000000002 ***
relevel(factor(ethnicity), ref = "White")Black  <0.0000000000000002 ***
relevel(factor(ethnicity), ref = "White")Hispanic <0.0000000000000002 ***

Zero-inflation model coefficients (binomial with logit link):
                                                Estimate Std. Error
(Intercept)                                       -7.941      0.317
relevel(factor(ethnicity), ref = "White")Black   -10.214     24.665
relevel(factor(ethnicity), ref = "White")Hispanic  -2.976      1.003
                                                z value
(Intercept)                                      -25.05
relevel(factor(ethnicity), ref = "White")Black    -0.41
relevel(factor(ethnicity), ref = "White")Hispanic  -2.97
                                                         Pr(>|z|)
(Intercept)                                     <0.0000000000000002 ***
relevel(factor(ethnicity), ref = "White")Black                0.679
relevel(factor(ethnicity), ref = "White")Hispanic             0.003 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 14
Log-likelihood: -9.37e+04 on 6 Df
```

Notice that the ouput produces two sets of coefficients. The first set of coefficients are used to model the count data. The fitted model is similar to the negative binomial and Poisson models.

$$\ln(\hat{\mu}_i) = -0.27 + 0.15(\text{Black}_i) + 0.22(\text{Hispanic}_i)$$

```
exp(c(0.15, 0.22)) - 1
```

```
[1] 0.162 0.246
```

- Blacks are stopped, on average, 16.2% more often than Whites with "stop and frisk" relative to the arrest rate from the previous year ($p < .001$).

- Hispanics are stopped, on average, 24.6% more often than Whites with "stop and frisk" relative to the arrest rate from the previous year ($p < .001$).

The second set of coefficients uses those same predictors to model the probability of a zero. In other words, they give the coefficients used to form a logistic regression model that give the log-odds of being a zero given the values of the predictors used in the model.

$$\ln\left[\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right] = -7.94 + -10.21(\text{Black}_i) - 2.98(\text{Hispanic}_i)$$

- The log-odds of having zero "stop and frisk" stops for Whites, relative to the arrest rate from the previous year, is -7.94.
- The log-odds of having zero "stop and frisk" stops for Blacks, relative to the arrest rate from the previous year, is -18.2.
- The log-odds of having zero "stop and frisk" stops for Hispanics, relative to the arrest rate from the previous year, is -10.9.

Translating these to probabilities, using

$$\hat{\pi}_i = \frac{\hat{\text{Odds}}}{1+\hat{\text{Odds}}}$$

```
exp(c(-7.94, -18.2, -10.9))
```

```
[1] 0.0003562065 0.0000000125 0.0000184582
```

```
exp(c(-7.94, -18.2, -10.9)) / (1 + exp(c(-7.94, -18.2, -10.9)))
```

```
[1] 0.0003560796 0.0000000125 0.0000184579
```

- The probability of having zero "stop and frisk" stops for Whites, relative to the arrest rate from the previous year, is 0.0003.
- The probability of having zero "stop and frisk" stops for Blacks, relative to the arrest rate from the previous year, is 0.0000000125.
- The probability of having zero "stop and frisk" stops for Hispanics, relative to the arrest rate from the previous year, is 0.0000184579

This suggests that Blacks and Hispanics are far less likely to have zero "stop and frisk" stops than Whites. And, even when there are "stop and frisk" stops, Black and Hispanics are stopped more often than Whites.