A Statistical Analysis of count data

MATH-493 - Applied Biostatistics: individual project

Federico Di Gennaro

Introduction

This statistical analysis examines the relationship between the apprentice migration between 1775 and 1799 to Edinburgh from 33 counties in Scotland. The study[1] provides the following four variables to try to explain the number of apprentice migrations:

- Distance: Distance from the county to Edinburgh.
- Population: Population (1000s) in the county.
- Degree_Urb: Degree of urbanization of the county (in %).
- Direction: Categorical variables that takes value in $\{1, 2, 3\}$ stays for: 1=North, 2=West, 3=South.

We define the outcome variable *Apprentices*, representing the number of apprentices originating from each of the 33 Scottish counties.
It is crucial to note that our outcome variable can be classified as *textitcountdata*, taking only discrete values. As such, it is important to acknowledge that our outcome will not follow a normal distribution (that is continuous), and therefore, traditional linear regression methods cannot be used. Given the nature of our data, we will need to explore alternative statistical methods that are more appropriate for count data analysis so that we can ensure that our results are both accurate and reliable.
From the literature otherwise, we know that Poisson regression is suitable to analyze count data as the ones of this analysis. Poisson regression is in the family of the so-called Generalized Linear Models (GLM). In a GLM, the response variable Y is related to the linear predictor, denoted by $\eta = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$, through a link function $g(\cdot)$. The link function establishes the relationship between the linear predictor and the expected value of the response variable, denoted by $\mathbb{E}[Y|x]$, such that $g(\mathbb{E}[Y|x]) = \eta$. In the context of Poisson regression (that relies on Poisson distribution), the link function is commonly defined as $g(x) = log(x)$.

Exploratory Data Analysis

Before fitting any kind of GLM, it is important to conduct exploratory data analysis. By doing so, we can gain insights into the relationships between variables and identify any potential issues or outliers in the data.
In this analysis, I am working with a dataset that consists of 33 observations, each corresponding to a different county in Scotland. For each observation, it was recorded the value for the dependent variable of interest, denoted by *Apprentices*, as well as the predictor variables that were introduced earlier in the report. Let's now have a look at the pair plot of the exogenous variables and at the distribution of the outcome.

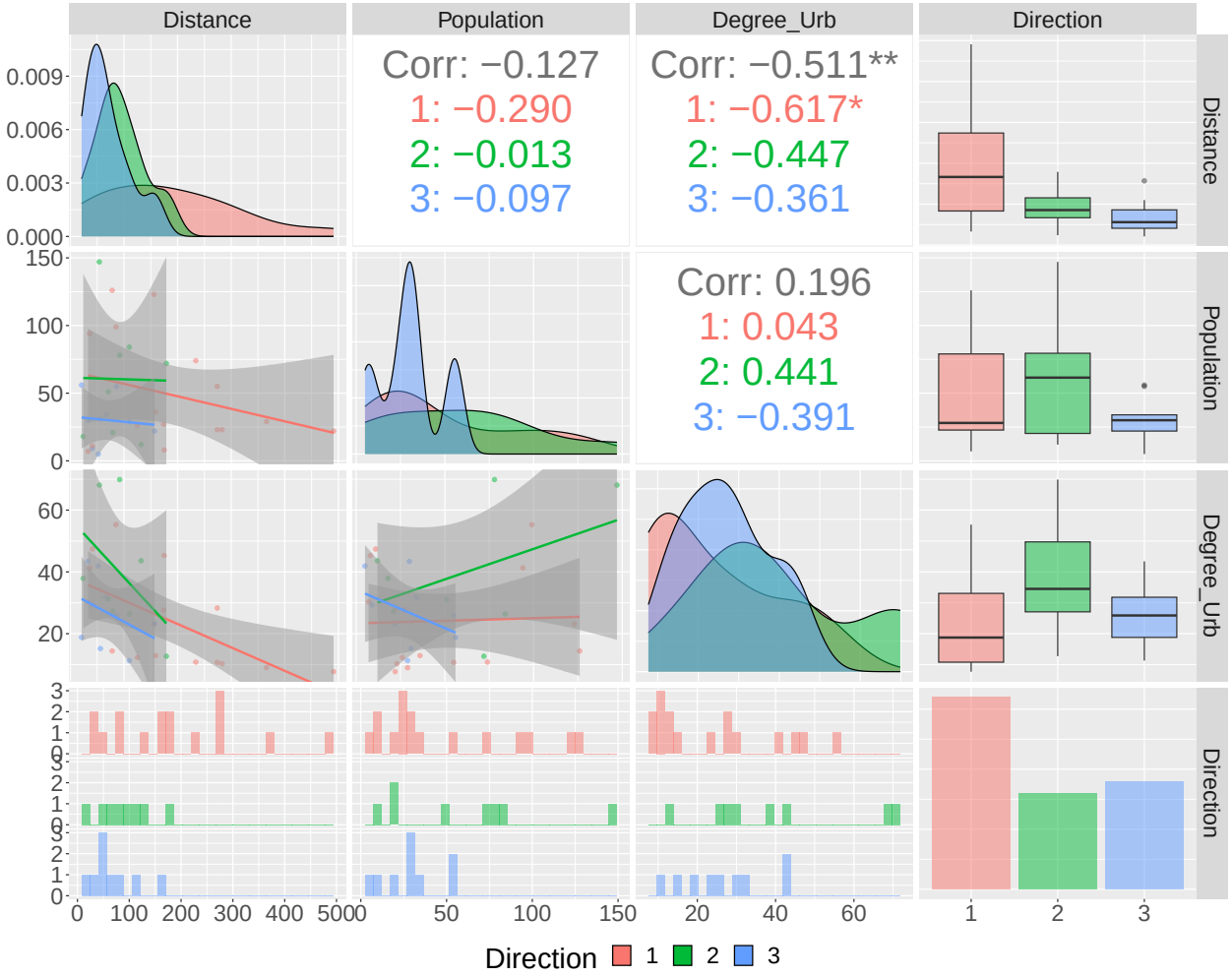

Figure 1: Pair lot of the regressors, divided by the factor variable Direction

Upon examining the plot, we can observe that the distributions of the variables *Distance*, *Population*, and *Degree_Urb* differ depending on the value of the factor variable *Direction*. *Distance* and *Degree_Urb* are the variables more correlated but from the decomposition of the plot above w.r.t. the factor variable *Direction* we can observe that for some value of *Direction*, also other pairs of variables are more correlated.

Additionally, it is worth noting that the variables *Distance* and *Population* exhibit a bit of skewness. Given that, it may be beneficial to apply a logarithmic transformation to these variables before fitting our Poisson regression models. This transformation can help to reduce the effect of extreme values on the results and improve the model's performance.

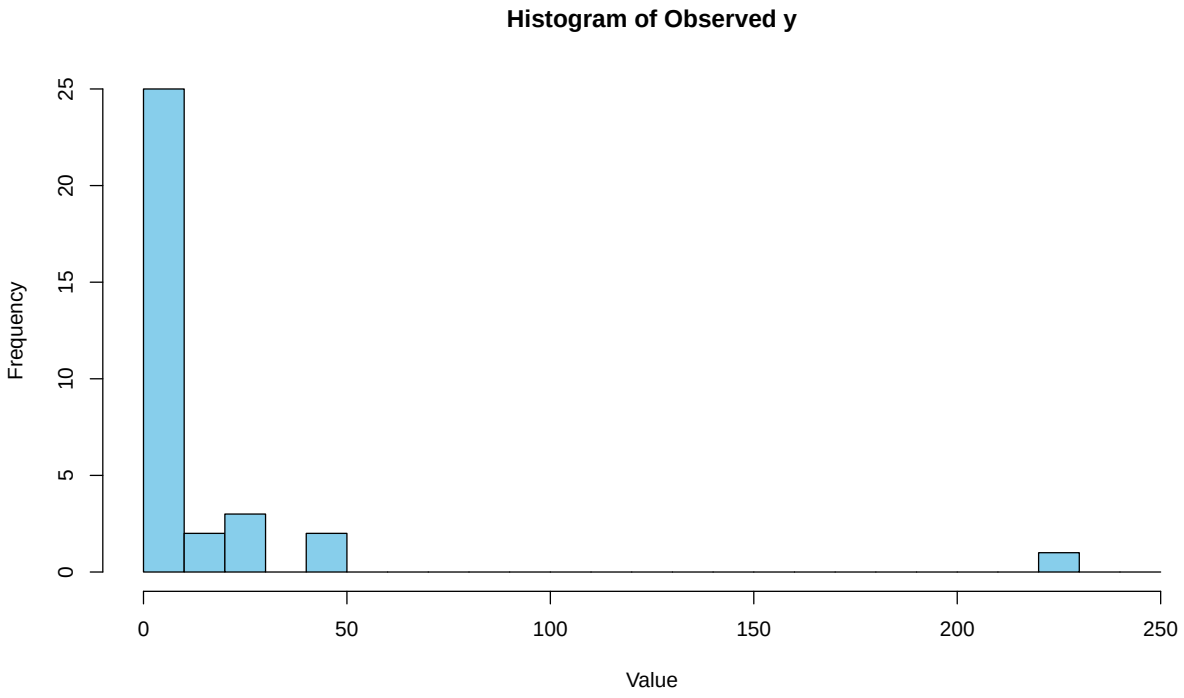

Figure 2: Histogram of outcome variable Apprentices

The histogram of the dependent variable *Apprentices* provides insight into its empirical distribution. Upon examination of the histogram, it is evident that there is an outlier observation that deviates significantly from the others. This outlier observation can lead to issues such as over-dispersion, which can affect the accuracy of the Poisson regression model.
It may be necessary to address this outlier observation before fitting the Poisson regression model to prevent potential issues with over-dispersion.

Model Fitting

Model highlights:

$Y = Apprentices;\ X_1 = Distance;\ X_2 = Population;\ X_3 = Degree\_Urb;\ X_4 = Direction.$

As already said in the introduction, we will use Poisson Regression to determine whether there is or not an effect between explanatory variables $X_i, i \in \{1, 2, 3, 4\}$ and the outcome variable Y.

Model 1: Poisson regression model.
Let's now define the first model in the study.

$$Model\ 1 : Y|X_1, ..., X_4 \sim Pois(\mu).$$
$$\mu = exp(\eta),\ \eta = \beta_0 + \beta_1 X_1 + ... + \beta_4 X_4.$$

The coefficients β_i of a Poisson Regression model are estimated using maximum likelihood estimation.

The Poisson regression model (cfr. Table 2 for a summary on its coefficients) in question appears to be afflicted with several issues. Firstly, as suggested in the exploratory data analysis, the model suffers from over-dispersion. To confirm this, we can fit a Quasi-Poisson model in R that outputs the estimate of the dispersion parameter (ϕ), which is found to be 7731.329. This value indicates an extremely large degree of over-dispersion, as the dispersion parameter should typically be close to 1 in a Poisson regression model.
It is essential to note that the significant coefficients in the original Poisson model may be misleading, as they are likely the result of the over-dispersion issue. The incorrect and artificially small standard errors that result from failing to adjust for over-dispersion can lead to artificially small p-values for the model coefficients.

Model 2: Base negative binomial model.
It is important to note that over-dispersion can have a significant impact on the model's results, as it can lead to incorrect conclusions about the relationships between the variables. Removing the outlier observation noticed in the EDA reduces the over-dispersion but does not remove it.
For this reason, it is better to use a model that takes it into account; this such a model can be a Quasi-Poisson model or a negative binomial model. The Quasi-Poisson model only provides a rough estimate of the dispersion parameter, so it may be necessary to consider alternative models, such as the Negative Binomial Regression model, which explicitly models over-dispersion[2].
For this reason from now on I will focus on fitting a negative binomial model. I chose a negative binomial model over a quasi-poisson model because when checking if the variance and the mean are proportional in the "mean-variance plot" (or "dispersion plot"), the points does not seems to be randomly scattered around a horizontal line at zero. Recall that the coefficients of a negative binomial model are implemented using maximum likelihood estimation.

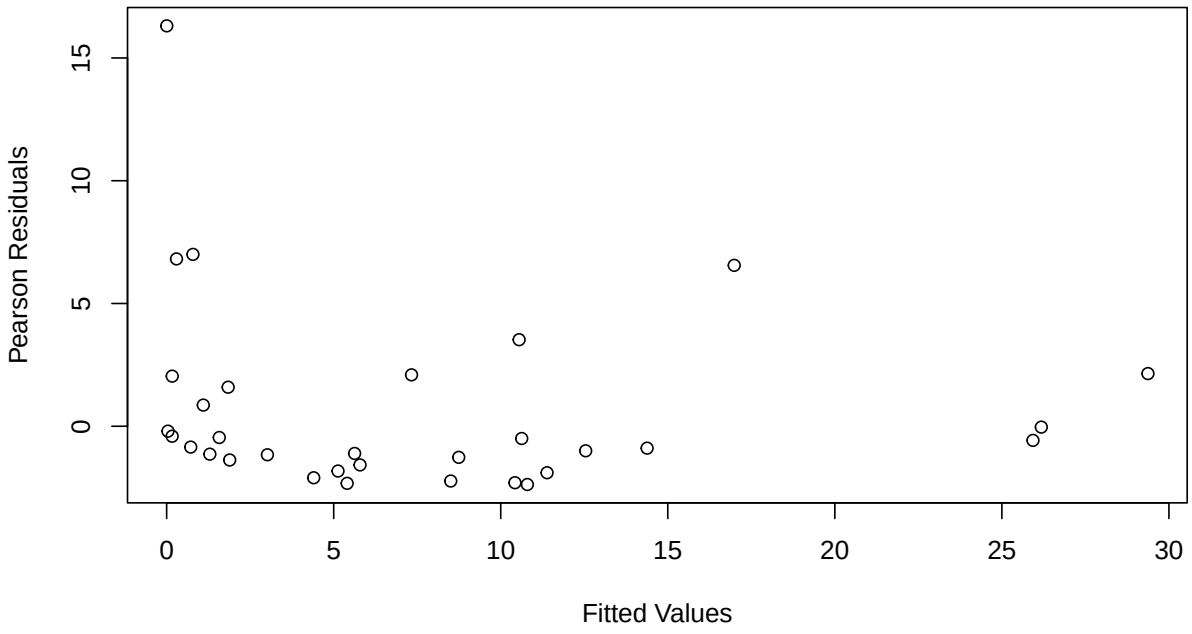

Figure 3: mean-variance plot

Given that the points form a curve that slopes downward from left to right, the variance increases slower than the mean.
The form of the model equation for negative binomial regression is the same as that for Poisson regression: the log of the outcome is predicted with a linear combination of the predictors.

$$\textit{Model 2}: Y|X_1, ..., X_4 \sim NegBin(\mu, \alpha).$$

$$\mu = exp(\eta),\ \eta = \beta_0 + \beta_1 X_1 + ... + \beta_4 X_4,\ \alpha \text{ is the over-dispersion parameter.}$$

For a summary on coefficients of this model, please see Table 3 in the Appendix.

Model 3: Negative binomial model with log-transformations.
As we previously observed in the exploratory data analysis, the variables *Distance* and *Population* exhibit skewness, which can cause issues when fitting a model. A common solution to this problem is to apply a logarithmic transformation to these variables. By doing so, we can better capture the relationship between these predictors and the response variable *Apprentices*. Please notice that the mathematical definition of *Model 3* is exactly the same as the one of *Model 2* with the exception of the *log* of the variables X_1 and X_2.

Let $X_1' = log(x_1)$, $X_2' = log(x_2)$, $X_3' = X_3$, $X_4' = X_4$; hence, we can define $Model3$ as:

$$\textit{Model 3}: Y|X_1', .., X_4' \sim NegBin(\mu, \alpha)$$

$$\mu = exp(\eta),\ \eta = \beta_0 + \beta_1 X_1' + ... + \beta_4 X_4',\ \alpha \text{ is the over-dispersion parameter.}$$

To assess the impact of this transformation, we compare a new model, denoted *Model 3*, to the original *Model 2*. While these models are not nested, we can still use the AIC as a metric for model selection. A lower AIC indicates a better model fit. In this case, *Model 3* has a lower AIC of 166.62 compared to *Model 2* with an AIC of 184.94. This indicates that *Model 3* is a better fit for our data and provides stronger statistical evidence for the relationship between our predictors and the response variable.

You can have a look at Table 4 in the Appendix for a summary on coefficients estimated for this model.

Model 4: Negative binomial model with log-transformations and interactions.
To improve the model's ability to explain the data, I have added complexity to *Model 3* by including interactions between all the exogenous variables. Let $Z_{i,j} = X_i' X_j'\ \forall i, j = 1, ...4$, $j > i$. Then we define $Model4$ as:

$$Model\ 4 : Y | X_1', .., X_4', Z_{1,2}, ..., Z_{3,4} \sim NegBin(\mu, \alpha).$$
$$\mu = exp(\eta),\ \alpha \text{ is the over-dispersion parameter.}$$

This time $\eta = \beta_0 + \beta_1 X_1' + ... + \beta_4 X_4' + \beta_{1,2} Z_{1,2} + ... + \beta_{3,4} Z_{3,4}$ You can have a look at Table 5 in the Appendix for a summary on coefficients estimated for this model.

To test whether *Model 4* is better than *Model 3*, we can use a likelihood ratio test.

$$\mathcal{H}_0 : \text{Model 3 is true.}$$

$$\mathcal{H}_1 : \text{Model 4 is true.}$$

#Df	LogLik	Df	Chisq	Pr(>Chisq)
7	-76.31	NA	NA	NA
16	-58.05	9	36.52	0

As we can see from the results above, the p-value of such a test is between 0.05 and 0.1 and so the result of the test crucially depends on the level of significance α we chose; the preferred model at a level 0.1 is *Model 4* but the preferred model at level 0.05 is *Model 3*. Given that the AIC is lower for *Model 3* (200 *vs* 166.62), we can argue that adding complexity does not improve too much the fitting. For this reason, the preferred model remains *Model 3*.

Final Model.
Starting from *Model 3*, I perform a stepwise selection based on the Akaike Information Criterion (AIC), a popular method to compare different models. In this way, we can balance the goodness-of-fit and the complexity of our model, and choose the one that best represents the underlying data-generating process.
After the selection, the variables keept are:

$$X_1' = log(Distance);\ X_2' = log(Population);\ X_3' = Degree\_Urb.$$

Hence, the final model fitted is:

$$Final\ Model: \hat{y} = exp(7.33 - 1.81X_1' + 0.83X_2' - 0.02X_3').$$

In this Final model, all the coefficients are significant at level 0.05 (cfr. Table 6 in Appendix). It is also good practice checking the correlation between the variables. We can do it by observing at the following plot.

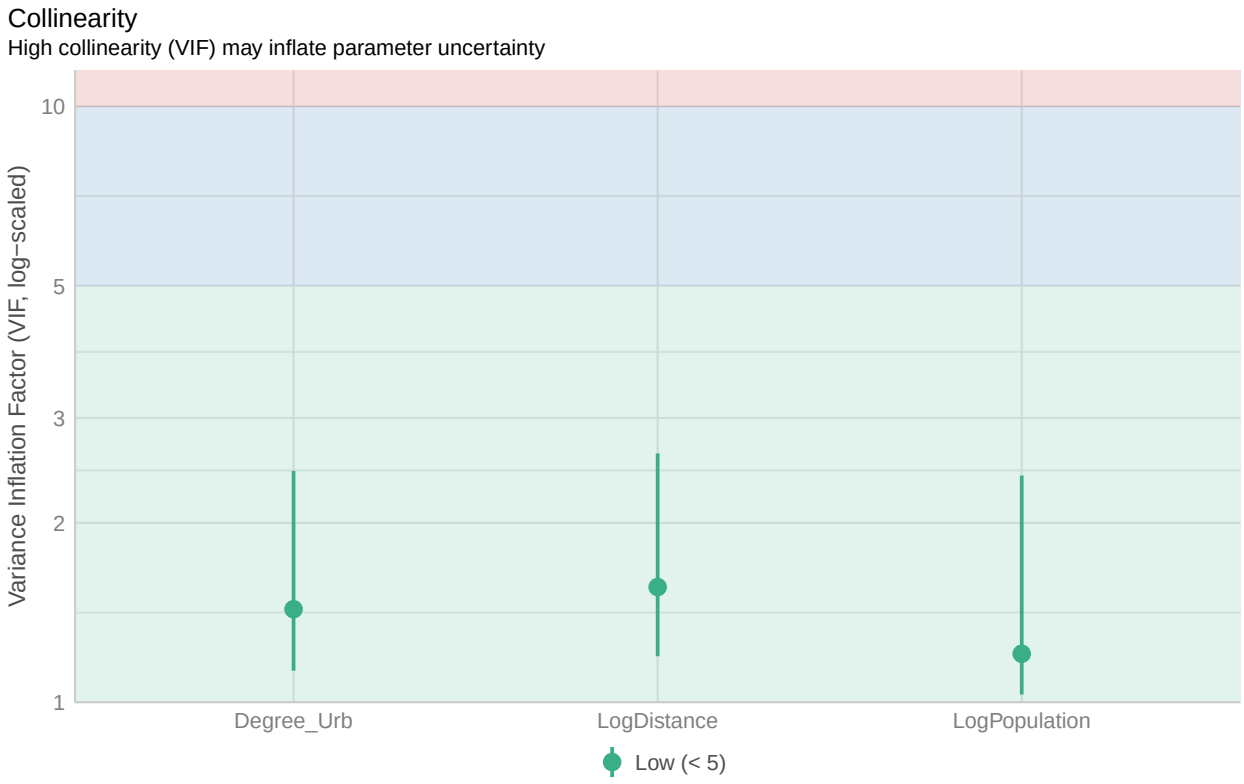

Figure 4: collinarity plot

From this plot I am happy to notice that the correlation among the variables is really low.

Model assessment

At this point, it is important to verify whether or not the assumptions of the *Final Model* (Negative Binomial Regression model) are satisfied. The assumption of a Negative Binomial Regression model are the following:

1) $Y|X_1', X_2', X_3' \sim NegBin(\mu, \alpha)$ where $\mu = exp(\hat{\beta}_0 + \hat{\beta}_1 X_1' + \hat{\beta}_2 X_2' + \hat{\beta}_3 X_3')$ and α is the over-dispersion parameter.
To do so, I use a diagnostic plot provided by the package *performance*.

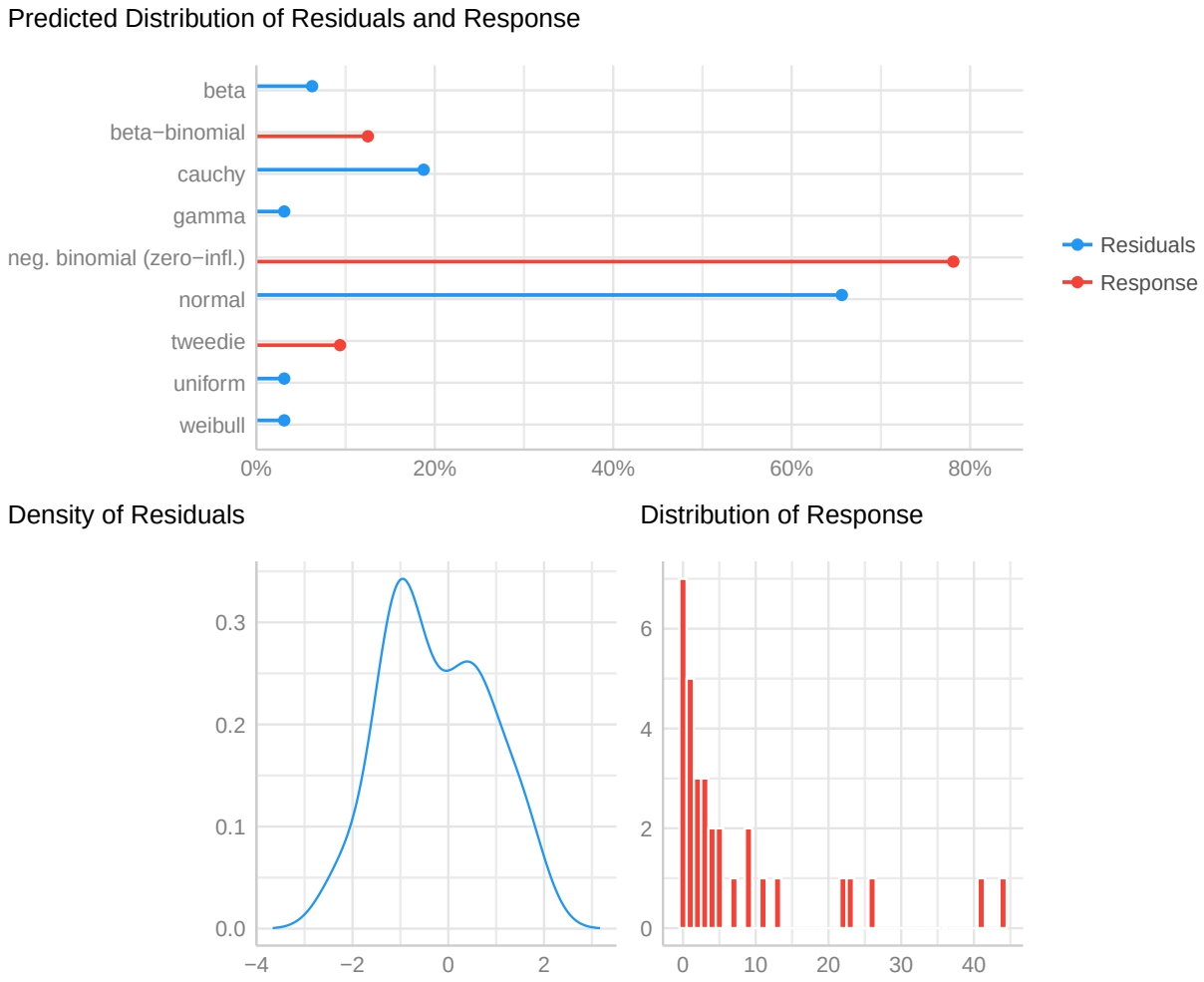

Figure 5: diagnostic plot for distribution

From the above plot, we can see that the distribution that better fits the data is a negative binomial. Using the usual euristic to check wheter or not it is better to fit a zero-inflated model, it turns out that there is no such a big evidence to do so.

2) The observations are independent.
There is no way to check this assumption because it depends on how the data were collected; we have to assume that it is true.

3) There is a linear relationship between log count and linear predictor.

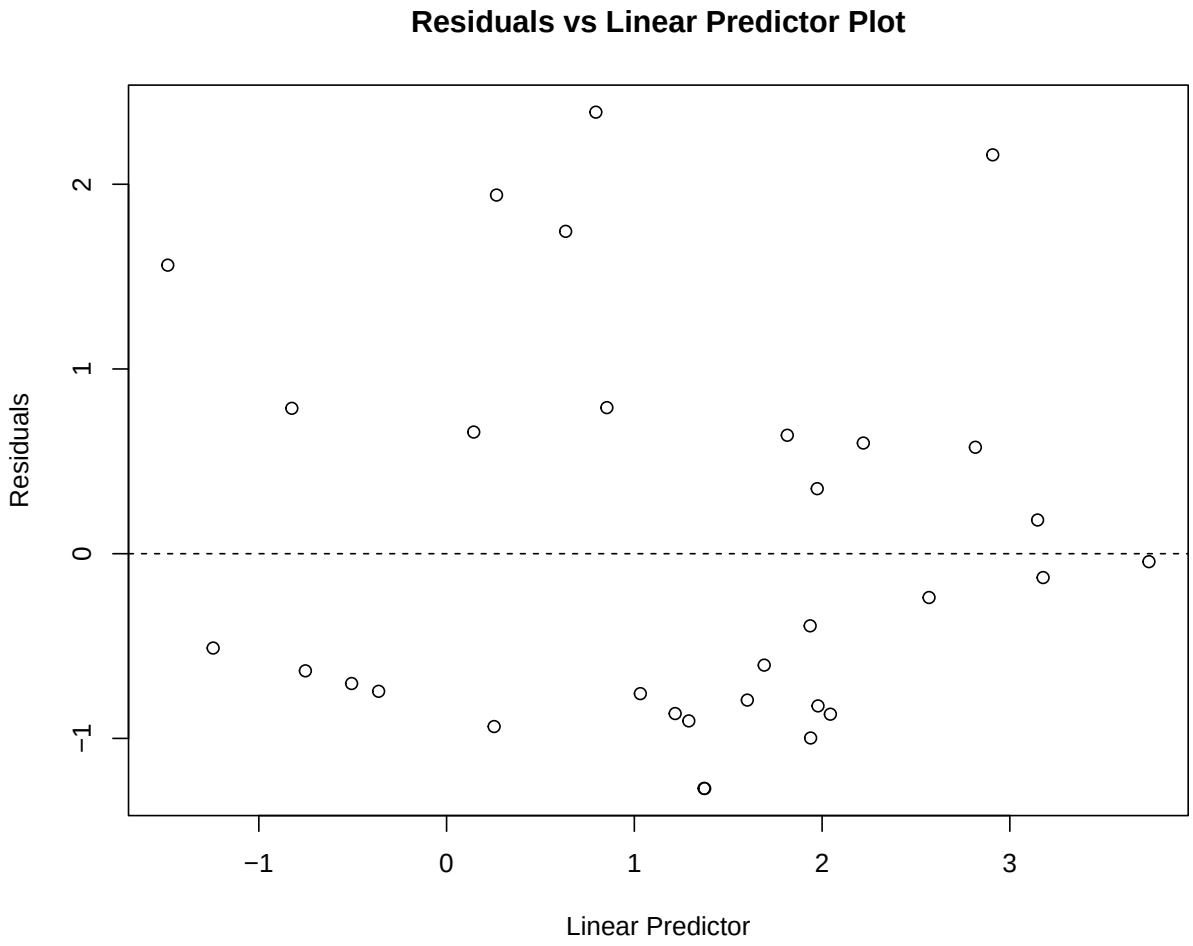

Figure 6: diagnostic plot of residuals vs fitted

The residuals are randomly scattered around the zero line and no evident pattern is shown in this diagnostic plot. For this reason, the assumption of a linear relationship between log count and linear predictor is met.

Conclusions

As we can see from the *Final Model*, the variables that play a role in determining the value of Apprentices for a given county are:
1) *log(Distance)*, with negative coefficient (-1.81). This means that the number of apprentice migrations decreases with a linear rate (due to the log in the definition of the variable) when the *log(Distance)* increases. This is what we would expect given that a long travel can decrease the willingness of an apprentice to migrate to Edinburgh.
2) *log(Population)*, with positive coefficient (0.83). This means that the number of Apprentices increases with a linear rate (due to the log in the definition of the variable) when the *log(Population)* increases.
Even this result is expected because when the population of a County is big, there is more competition and people are willing to travel more to get better opportunities.
3) Finally, the last variable (*Degree_Urb*), has a negative coefficient (-0.02) that is very small (but from Table 6 we can see that it is significantly different from zero). This means that the number of apprentices migration decreases when the *Degree_Urb* increases, but due to the magnitude of its β the influence of this variable on the outcome variable is not marked.

Please notice that the direction is not a significant variable, meaning that probably the things that affect most apprentices migrations are more about the social environment that the city you are offered (*Degree_Urb*, *Population*) instead of the direction of it.

References

[1] A. Lovett and R. Flowerdew (1989). "Analysis of Count Data Using Poisson Regression", Professional Geographer, 41, pp. 190-198.
[2] Michael L. Zwilling (2013). "Negative Binomial Regression", Mathematical Journal.

Appendix

Model 1:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	4.25	0.25	17.16	2e-16
Distance	-0.03	1.93e-03	-17.59	2e-16
Population	0.02	1.52e-03	14.01	2e-16
Degree_Urb	-0.04	4e-03	-8.84	2e-16
Direction2	0.23	0.18	1.27	0.21
Direction3	1.11	0.15	7.38	1.62e-13

Table 2: Summary of *Model 1* coefficients

Model 2:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.45	0.83	2.95	3.13e-03
Distance	-0.01	3.22e-03	-3.19	1.39e-03
Population	0.01	5.23e-03	2.66	7.85e-03
Degree_Urb	-8.01e-03	1.44e-02	-0.56	0.58
Direction2	-0.33	0.50	-0.66	0.51
Direction3	0.27	0.53	0.51	0.61

Table 3: Summary of *Model 2* coefficients

Model 3:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	6.93	1.35	5.14	2.78e-07
LogDistance	-1.79	0.27	-6.62	3.57e-11
LogPopulation	0.89	0.17	5.16	2.47e-07
Degree_Urb	-0.02	0.01	-1.44	0.15
Direction2	-0.49	0.37	-1.35	0.18
Direction3	0.10	0.35	0.30	0.77

Table 4: Summary of *Model 3* coefficients

Model 4:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.07	5.47	-0.56	0.57
LogDistance	0.21	0.98	0.21	0.83
LogPopulation	2.86	1.18	2.42	0.02
Degree_Urb	0.05	0.10	0.54	0.59
Direction2	5.56	2.46	2.26	0.02
Direction3	7.86	3.29	2.39	0.02
LogDistance:LogPopulation	-0.37	0.23	-1.61	0.10
LogDistance:Degree_Urb	-9.6e-03	0.01	-0.65	0.52
LogDistance:Direction2	-2.16	0.9	-2.39	0.02
LogDistance:Direction3	-1.57	0.62	-2.55	0.10
LogPopulation:Degree_Urb	-8.85e-03	0.01	0.718	0.47
LogPopulation:Direction2	1.12	0.72	1.56	0.12
LogPopulation:Direction3	-0.49	0.26	-1.86	0.06
Degree_Urb:Direction2	-0.03	0.03	-1.00	0.32
Degree_Urb:Direction3	0.02	0.03	0.56	0.58

Table 5: Summary of *Model 4* coefficients

Final Model:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	7.33	1.27	5.77	7.88e-09
LogDistance	-1.81	0.27	-6.64	3.01e-11
LogPopulation	0.83	0.17	4.74	2.19e-06
Degree_Urb	-0.02	0.01	-2.01	0.04

Table 6: Summary of *Final Model* coefficients