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 regions in Scotland. The study utilizes the following four variables to try to predict the number of apprentices migration:

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

The outcome variable, as said before, is the number apprentices migration that in this analysis we called "Apprentices". It is immediately important to notice the variables we are working with; in particular, we can notice that the outcome variable "Apprentices" can be seen as "count data", taking only discrete values. For that reason, we will not end up with an outcome normally distributed and so linear regression cannot be used.

From literature otherwise, we know that Poisson regression is suitable to analyse count data (in this case our Y is a the number of apprentices, i.e. it can be seen as count data). Poisson regression is in the family of the so-called Generalized Linear Models (GLM).
Generalized Linear Models are models in which response variables follow a distribution other than the normal distribution. In GLMs, the response variable is connected to the linear predictor $\eta = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$ by a link function $g(\cdot)$, that describe the functional relationship between η and the mathematical expectation of the response variable: $g(\mathbb{E}[Y|x]) = \eta$.
For Poisson Regression, $g(x) = log(x)$.

Poisson regression relies on Poisson distribution; we say that a discrete random variable X is distributed as a Poisson with parameter λ ($X \sim Poisson(\lambda)$) if it has the following density function:

$$\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Exploratory Data Analysis

Before starting in fitting the model, it is a good practice to explore the available data. In this way, we can already notice some insights about what the model will tell us and in which way each variable can effect the model itself. The dataset provided has 33 rows, corresponding to the different counties of Scotland under study, and for each row we have the dependent variable under study (Apprentices), and the registered values for the regressors mentioned in the introduction.

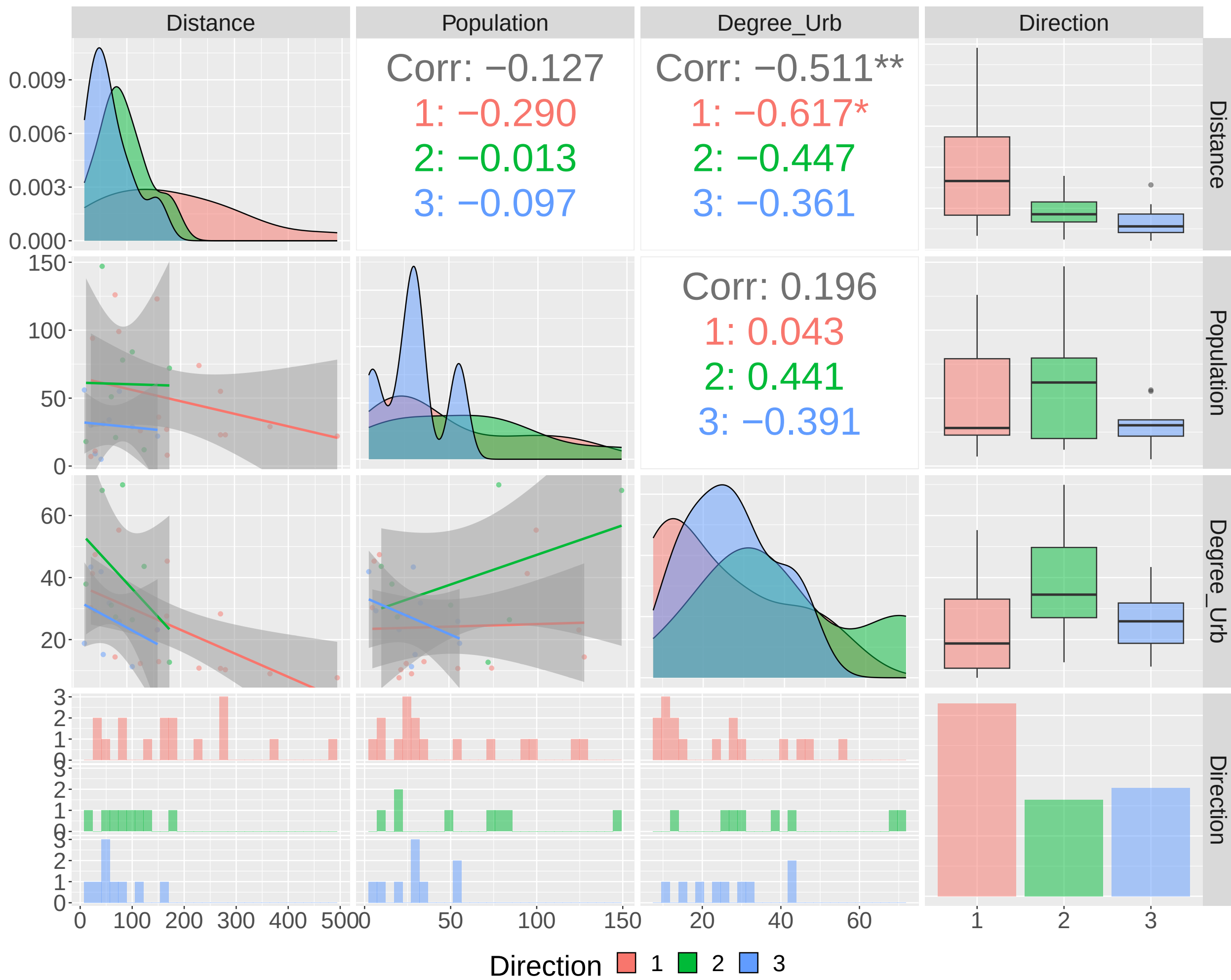

Figure 1: Pairs Plot of the explanatory variables. Their distribution is divided with respect to the value of the variable Direction (factor)

From this plot, we can notice that the three variables *Distance*, *Population* and *Degree Urb* change distribution with respect to the value of the value of the factor variable *Direction*. Please notice also that the variables *Distance* and *Population* are really skewed; for this reason it can be worth at some point of the analysis trying to apply a logarithmic transformation to them in fitting our models.

With the histogram of the outcome variable *Apprentices*, I can observe the shape of the

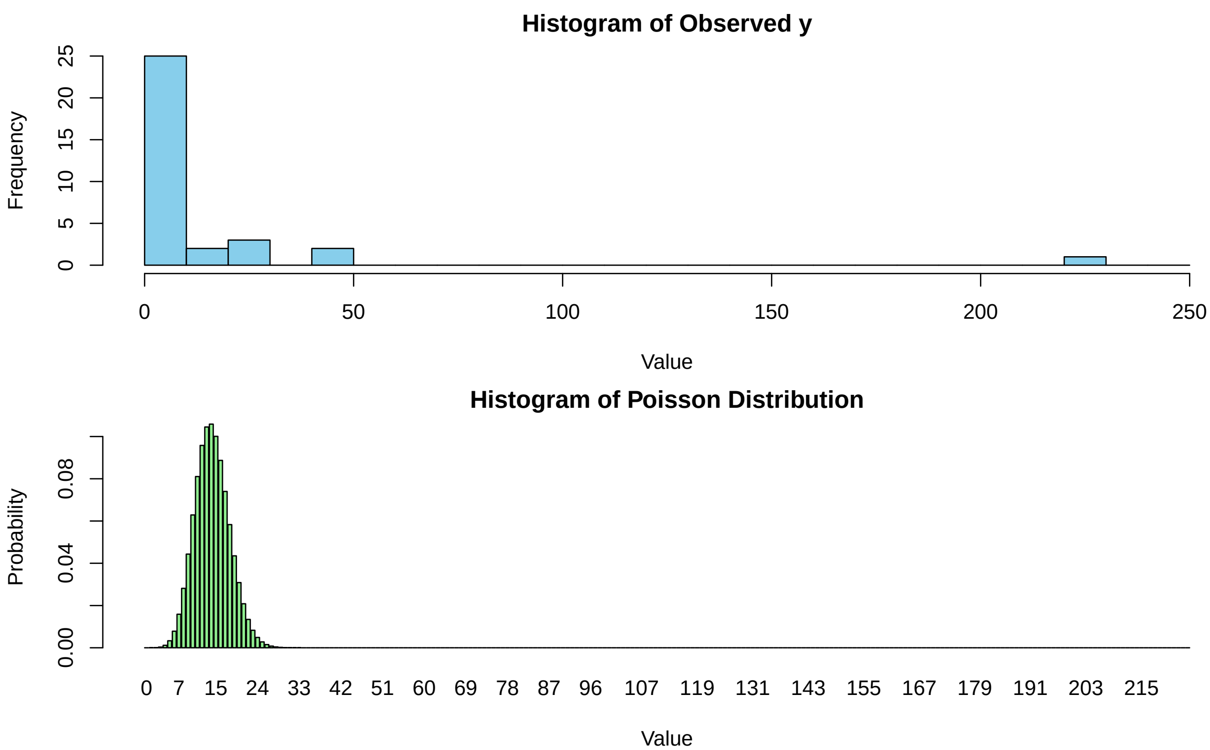

Figure 2: Histogram of outcome variable Apprentices

empirical distribution of that variable. Specifically, it is already noticeable that there is an observation that is really far away from the others; this can leads problems such as 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 effect between explanatory variables $X_i, i \in \{1, 2, 3, 4\}$ and the outcome variable Y.

Model 1.
Let's now define the first model in the study. As already said before, let

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

Hence, the model is defined as

$$log(\mathbb{E}[Y|x]) = \eta$$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	4.25	0.25	17.16	0.00
Distance	-0.03	0.00	-17.59	0.00
Population	0.02	0.00	14.01	0.00
Degree_Urb	-0.04	0.00	-8.84	0.00
Direction2	0.23	0.18	1.27	0.21
Direction3	1.11	0.15	7.38	0.00

This model seems to have a lot of problems. Firstly, as already guessed in the exploratory data analysis, the model suffers of over-dispersion. The residual deviance (256.31) is bigger than the degrees of freedom (28). For this reason it seems that there is a larger variance in observed counts than expected from the Poisson assumption, meaning that there is over-dispersion and that the Poisson model does not fit well.
Another way to observe the over-dispersion, is fitting a quasi-Poisson model: the output in R will have the same coefficient as *Model 1* but from this new model we can extrapolate the real dispersion parameter (ϕ) of the data that turns out to be 7731.329 (i.e. there is an incredibly huge over-dispersion because that parameter should be around 1 in a Poisson Regression model). Please observe that we should not be fooled by the super significant coefficients. Those extremely low p-values are exactly the consequences of the over-dispersion issue. Without adjusting for over-dispersion in fact, we use incorrect and artificially small standard errors leading to artificially small p-values for model coefficients.

Model 2. The problem of over-dispersion, can be caused by the extremely outlier observation of variable *Apprentices* that we already observed in the EDA. For this reason I am removing this observation from my data and then try to fit again the same model stated for *Model 1*. Even removing the strange observation, the model still suffers of over-dispersion ($\mu = 7.59$ and $\sigma^2 = 131.86$). In this case otherwise, the dispersion parameter is no longer as large as before ($\phi = 18.43$) and for this reason it is worth trying to work with this data.
Due to the fact that over-dispersion is still present, we should use a Quasi-Poisson model; this model would relax the assumption of $\phi = 1$. In this new model, the additional over-dispersion parameter $\phi = 18.43$ might capture a large amount of the variation of the response variable. Note that since the underlying data is the same, the residuals deviance is the same. For this reason, from now on I will work with Quasi-Poisson models and I will try to see if adding complexity leads to models that explain better the data.
(Residual deviance of Model 1 (on 27 degrees of freedom) is equal to 256.31).

Model 3. As said in the previous exploratory data analysis, it is worth trying to transform the explanatory variables *Distance* and *Population* with a logarithmic transformation. This can avoid problems coming from the skewednees these variables.
I test now whether or not there is statistical evidence that this new model is significantly better than Model 2. To do so, I use a ????? test.

$$H_0 : \text{Model 2 is sufficient}$$

It is good to notice that the over-dispersion parameter ϕ is decreasing too (now it is equal to 3.37). Let's see if also adding interactions between the variables make ϕ even lower.

Model 4. Now I am adding at the Quasi-Poisson regression model in *Model 3* the interactions between the whole variables. So I am keeping the logarithm in the variable and I am adding the interactions.
Given that the model it is already very complex, I decide to not add further transformations; in this way we have a model that is yes complex, but it is still interpretable in its variables.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-3.07	6.61	-0.46	0.65
LogDistance	0.21	1.19	0.18	0.86
LogPopulation	2.86	1.43	2.00	0.06
Degree_Urb	0.05	0.12	0.44	0.66
Direction2	5.56	2.98	1.87	0.08
Direction3	7.86	3.97	1.98	0.06
LogDistance:LogPopulation	-0.37	0.28	-1.33	0.20
LogDistance:Degree_Urb	-0.01	0.02	-0.54	0.60
LogDistance:Direction2	-2.16	1.09	-1.98	0.06
LogDistance:Direction3	-1.57	0.74	-2.11	0.05
LogPopulation:Degree_Urb	-0.01	0.01	-0.59	0.56
LogPopulation:Direction2	1.12	0.87	1.29	0.21
LogPopulation:Direction3	-0.49	0.32	-1.54	0.14
Degree_Urb:Direction2	-0.03	0.03	-0.83	0.42
Degree_Urb:Direction3	0.02	0.03	0.46	0.65

As already said before, in this table the p-values are not erroneous and we can notice that there are five variables that are significant at level $\alpha = 0.05$.
In order to see if this new model is statistically significant better than Model 3, I use an F-test [REFERENCE Venables and Ripley Modern Applied Statistics in S].
Drop-in-deviance tests in fact, can be adjusted for overdispersion in the quasi-Poisson model. In this case, we can divide the test statistic (per degree of freedom) by the estimated dispersion parameter and compare the result to an F-distribution with the difference in the model degrees of freedom for the numerator and the degrees of freedom for the larger model in the denominator.

$$H_0 : \text{The reduced model (Model 3) is sufficient}$$

$$F = \frac{\frac{\text{drop in deviance}}{\text{difference in df}}}{\phi}(statistic)$$

Resid. Df	Resid. Dev	Df	Deviance	F	Pr(>F)
26	79.59	NA	NA	NA	NA
17	29.01	9	50.58	3.85	0.01

As we can see from the anova table above, the p-value of such a test is really small; for this reason the preferred model is Model 4.

Model assessment

At this point, it is important to verify whether or not the assumptions of the Quasi-Poisson model preferred so far (Model 4) are satisfied. The assumptions of a Quasi-Poisson model are the following:
1) There is a linear relationship between the logarithm of the frequency or rate and equal increment changes in the explanatory variable.
3) Errors are independent of each other.

Please notice that in the Quasi-Poisson regression we have relaxed the assumption that the variance is equal to the mean.

In order to check these assumptions, we can look at the following plots.

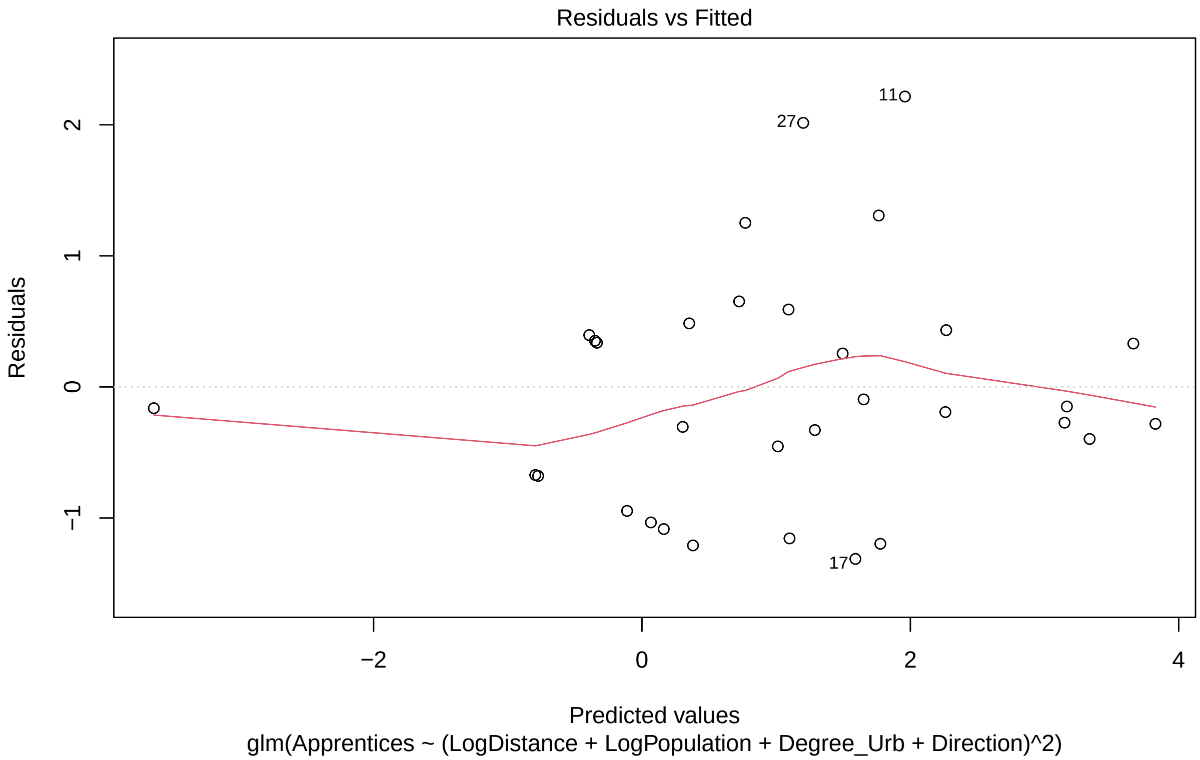

Figure 3: Checking model assumptions

Furthermore, it is a good practice for this model to check other possible issues such as over-dispersion, outlier presence, collinearity, etc…

From the plots above, we can see that the model is over-dispersive, and this makes our intuition of fitting a Quasi-Poisson regression model right and we can also notice that the data does not present outliers. Finally, it is important to notice that the plot of multicollinearity was made without the interaction variables: if the model contains interaction variables in fact, VIFs might be inflated.

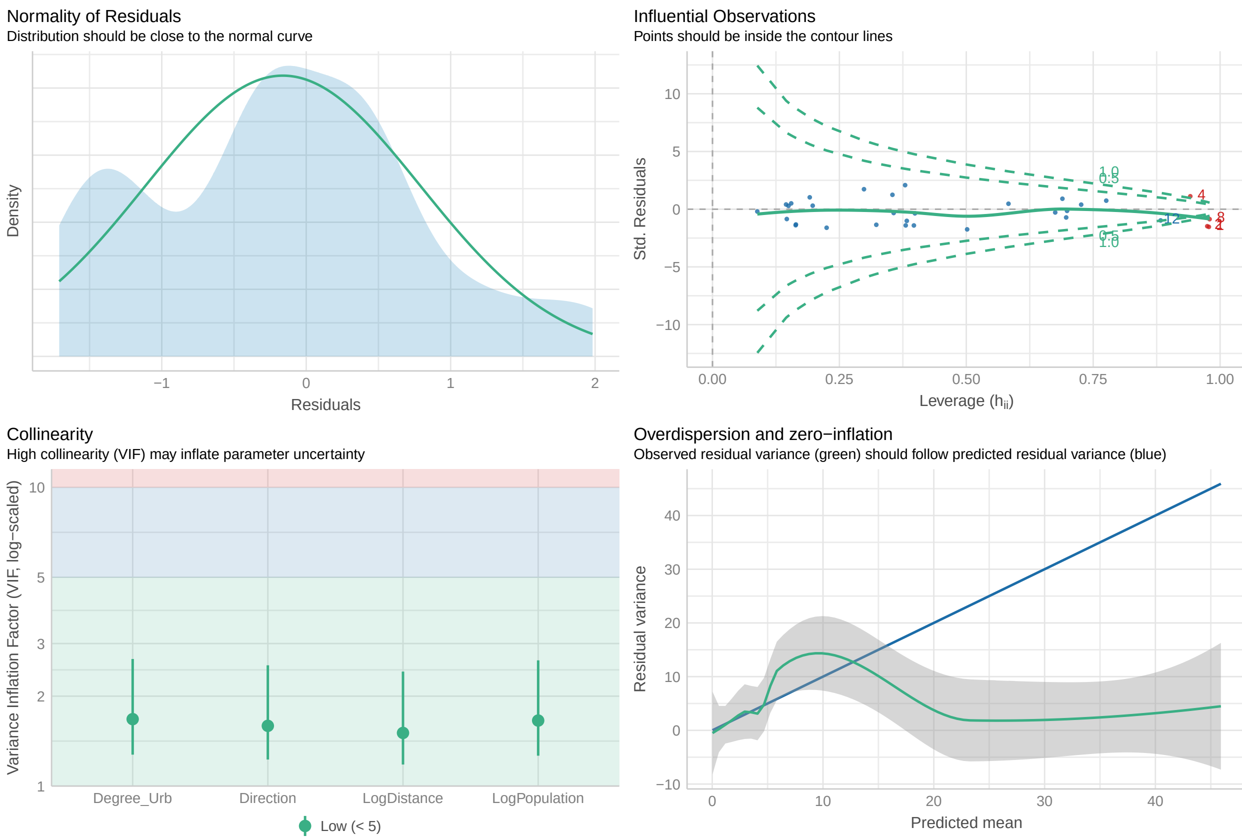

Figure 4: Checking further issues

```
## # Check for zero-inflation
##
##    Observed zeros: 7
##   Predicted zeros: 6
##             Ratio: 0.86
```

Final model and conclusion

From the analysis above, the final model is a Quasi-Poisson model $$