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.

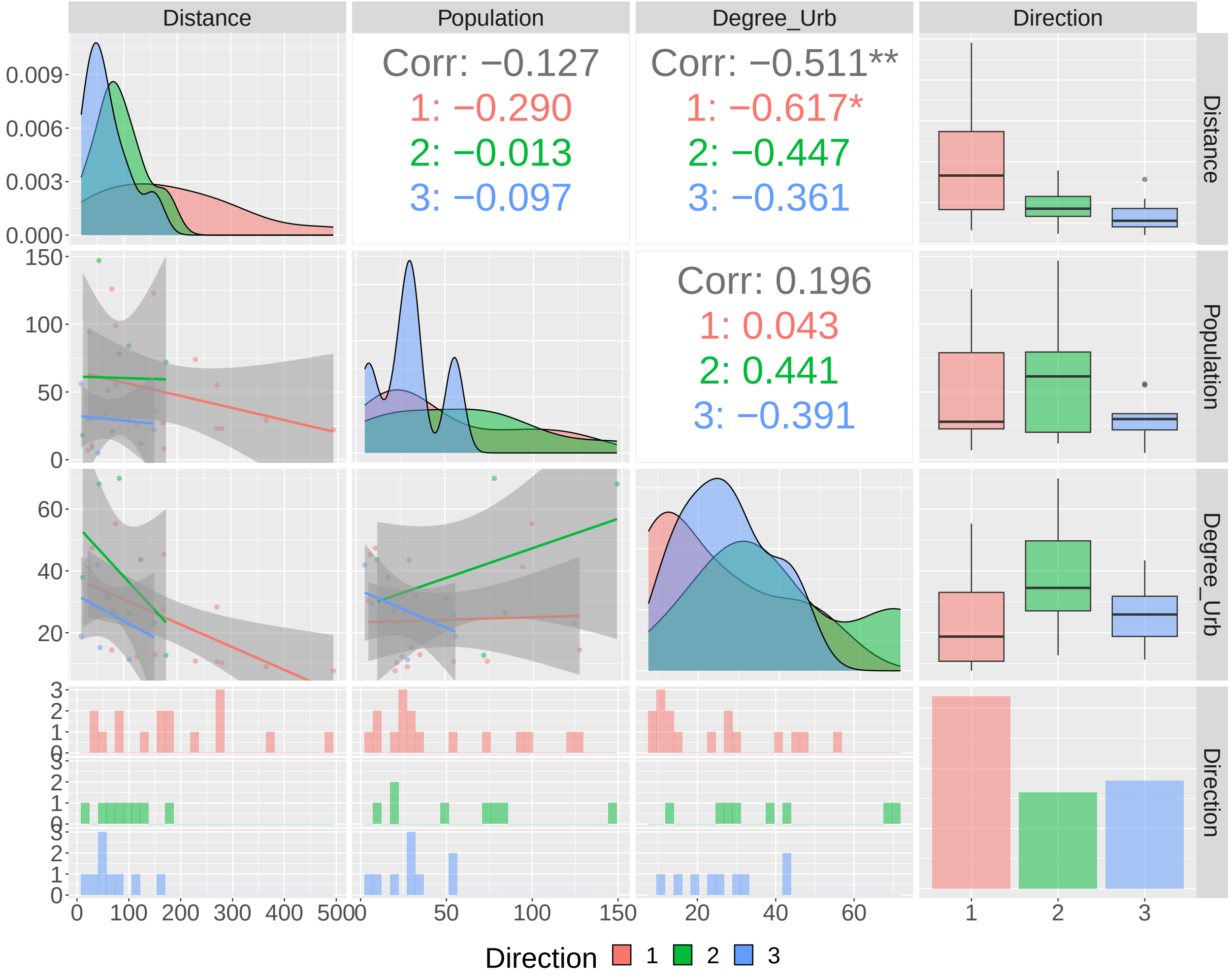

Figure 1: Pairs Plot of the explanatory variables. Their distribution is divided with respect to the value of the variable Direction (factor)

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

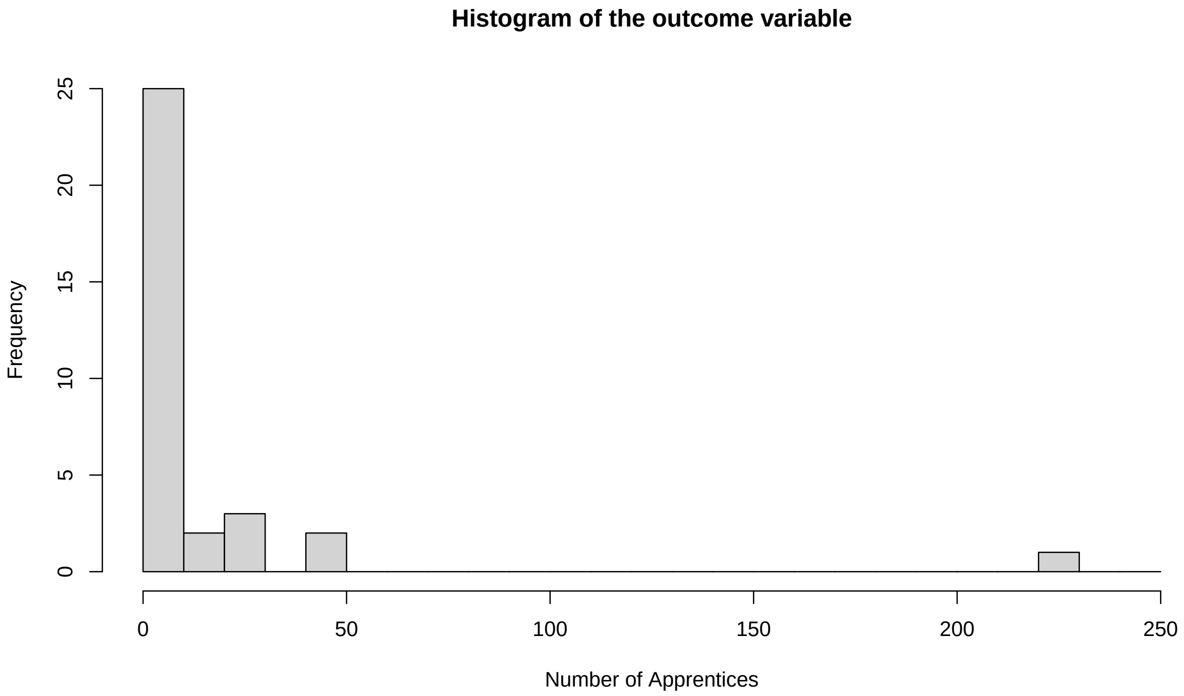

Figure 2: Histogram of outcome variable Apprentices

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:

Let

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

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 problems seems to have a lot of problem. 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 a huge over-dispersion because that parameter should be around 1 in a Poisson Regression model). This problem can be related to the extremely un-related observation of variable *Apprentices* that we observed in the EDA. For this reason I am removing this observation from my data and then try to fit again the same model.

Model 2. As already said above, this model is the same model as *Model 1* where I simply removed the outlier observation that probably made *Model 1* over-dispersive.
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 over-dispersion, modeling this data using the Poisson family assumption yield erroneous p-values regarding model variables. For this reason I decide to use a quasi-Poisson model, that 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.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.25	1.40	2.33	0.03
Distance	-0.02	0.01	-2.07	0.05
Population	0.01	0.01	1.66	0.11
Degree_Urb	-0.01	0.02	-0.67	0.51
Direction	0.10	0.41	0.24	0.82

Under these assumptions we can see that the parameters significant at a level $\alpha = 0.05$ are β_0(intercept), β_1 (coefficient of the variable *Distance*) and β_2 (coefficient of the variable *Population*).
The significant variable *Distance* as a negative sign in its parameter, meaning that at an

increase of *Distance* corresponds a decrease of the number of *Apprentices*. The significant variable *Population* as a positive sign in its parameter, meaning that at an increase of *Distance* corresponds an increase of the number of *Apprentices*. From now on, I will consider this model (Quasi-Poisson regression model) as **base model**.

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.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.20	1.20	6.01	0.00
LogDistance	-1.99	0.28	-7.19	0.00
LogPopulation	0.94	0.18	5.11	0.00
Degree_Urb	-0.02	0.01	-2.06	0.05
Direction	0.20	0.17	1.20	0.24

From these model, we can observe that the variables *Distance* and *Population* are even more significant then in *Model 3*, and their interpretation does not change because the sign of their parameters is still the ones as in *Model 3*. This problem is better than *Model 3* MAKE THE STATISTICAL TEST!!!!!!! It is also good to notice that the over-dispersion parameter ϕ is decreasing too (now it is equal to 3.37)

Model 4. Now I am adding at the Quasi-Poisson model *Model 4* the interactions between the whole variables.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-5.73	5.36	-1.07	0.30
LogDistance	0.78	1.01	0.77	0.45
LogPopulation	2.67	1.01	2.65	0.02
Degree_Urb	0.03	0.10	0.35	0.73
Direction	3.77	1.47	2.57	0.02
LogDistance:LogPopulation	-0.29	0.22	-1.33	0.20
LogDistance:Degree_Urb	-0.02	0.02	-1.01	0.32
LogDistance:Direction	-0.73	0.30	-2.42	0.02
LogPopulation:Degree_Urb	0.00	0.01	-0.10	0.92
LogPopulation:Direction	-0.27	0.15	-1.77	0.09
Degree_Urb:Direction	0.01	0.01	0.95	0.35

From the table above, we can observe that there are several variables that are significant at level $\alpha = 0.05$ but a lot of them are not. For this model, I can apply a step-wise backward elimination procedure to determine the best subset of independent variables that explain the variation in the dependent variable.