Poisson Regression Analysis of Apprentice Migration

Nino Gerber

2025-05-17

Introduction

This report investigates factors influencing the number of apprentices migrating from various regions to Edinburgh. The response variable is the **count of apprentices**, and the predictors are **distance from Edinburgh**, **population of the region**, **degree of urbanization**, and **direction from Edinburgh**. The scientific question is: *Which geographical and demographic factors significantly predict the number of apprentices migrating to Edinburgh between 1775 and 1799, and how do these relationships vary across regions?*

The dataset records the number of apprentices moving to Edinburgh between 1775 and 1799 from other Scottish counties. During this period, Edinburgh was a significant center for trade and education, attracting young individuals looking for apprenticeship opportunities. Understanding the patterns of apprentice migration during this time provides valuable insights into the socio-economic factors influencing labor mobility in the 18th-century in Scotland. (Lovett & Flowerdew, 1989).

Descriptive Statistics

This section provides an overview of the key variables in the dataset. The dataset used in this analysis was obtained from http://users.stat.ufl.edu/~winner/data/apprentice.txt. It contains the following variables:

- `region`: Region name
- `apprentices`: Number of apprentices (response variable)
- `distance`: Distance from Edinburgh (in miles)
- `population`: Population of the region (in thousands)
- `urban`: Urbanization score (numeric)
- `direction`: Cardinal direction from Edinburgh (coded as a factor: 1=North, 2=West, 3=South).

Numerical Variables: Table 1 presents the minimum, median, mean, and maximum values for four continuous variables:

Table 1: Clean Summary Statistics of the Main Variables

Variable	Min	Median	Mean	Max
Distance	21.0	92.0	131.8	491.0
Apprentices	0.0	3.0	14.2	225.0
Population	5.0	30.0	46.6	147.0
Urbanization	7.7	27.3	28.6	69.9

The figures in table 1 show that while most counties contributed few apprentices (median = 3), some—such as Midlothian—had significantly higher counts, driving the mean up to 14.2. Similarly, distances vary widely, with some regions more than 400 miles from Edinburgh.

Categorical Variable, Direction: The dataset also categorizes each county into one of three cardinal directions relative to Edinburgh: North, West, and South. The majority of counties are located in the North (49%), followed by the South (27%) and West (24%). There are 33 different counties in the data set and therefore 33 records.

Counties with the Most and Fewest Apprentices

To better understand spatial disparities in apprentice migration, table 2 highlights the counties with the highest number of apprentices

Table 2: Top 5 Counties with the Most Apprentices

region	apprentices	distance	population	urban	direction
Midlothian	225	21	56	18.8	South
East Lothian	44	33	30	43.4	South
Fife	41	36	94	41.3	North
Perth	26	78	126	14.4	North
Lanark	23	54	147	68.1	West

These top counties are either located close to Edinburgh (e.g., Midlothian, East Lothian) or have high population and urbanization scores, suggesting accessibility and economic opportunity may have facilitated apprentice movement

By investigating the dataset, we observe that 7 out of 33 Scottish counties report zero apprentices. This is likely to impact the modeling later on and should be taken into account. These counties tend to have low population sizes (ranging from 8 to 29 thousand) and high distances from Edinburgh (between 79 and 366 miles). Their urbanization scores vary from 9 to 43.6. This suggests that accessibility—captured by distance—as well as population size and urbanization, may influence apprentice migration. Notably, the counties with zero apprentices are evenly distributed across the North, South, and West.

Exploratory Data Analysis

This section provides an overview of the data using correlation analysis, histograms, and scatterplots to guide model specification.

Correlation Matrix and Distribution Overview

To understand both the individual variables and their pairwise relationships, we use a correlation matrix with embedded histograms and scatterplots. This helps guide model choice and transformations.

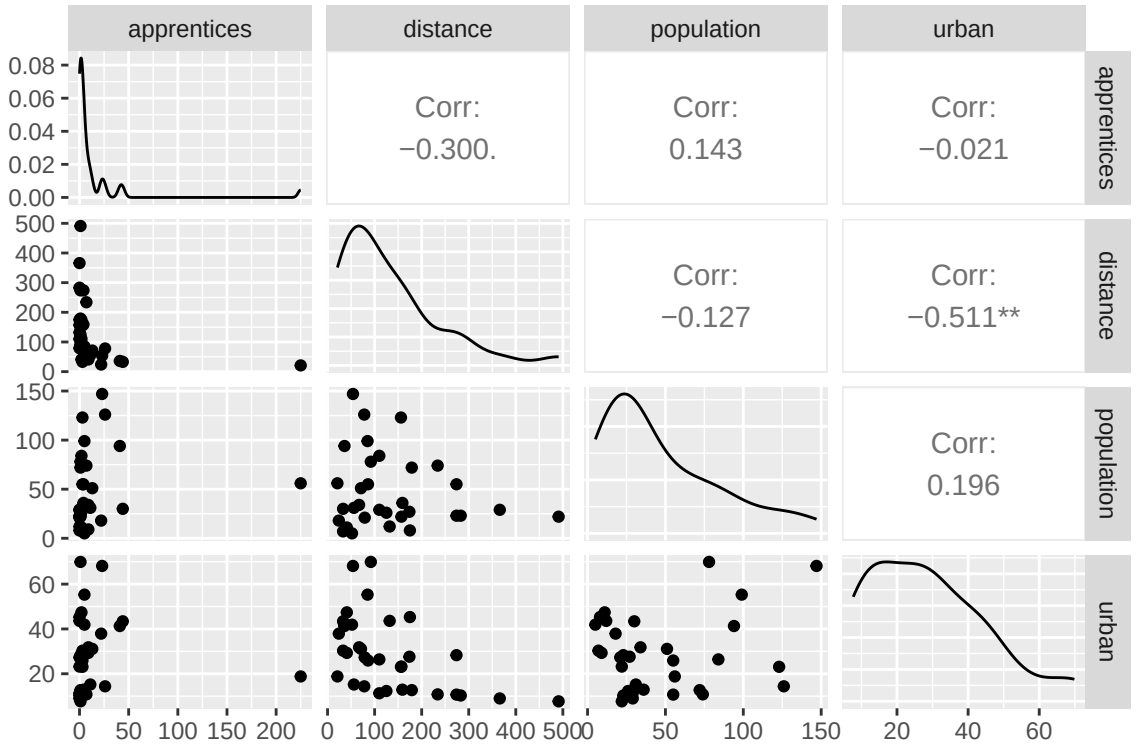

Figure 1: Correlation Matrix and Distribution Plots

Apprentices (Response Variable): Apprenticeship counts are heavily right-skewed, with most regions sending few apprentices and one clear outlier (Midlothian).

Distance (Predictor): Shows a moderate negative correlation with apprentices (r = –0.30), making it a key predictor. Distances range from 21 to 491 miles, with most counties located within 50–150 miles of Edinburgh.

Population (Predictor): The distribution is right-skewed, with most counties having fewer than 50,000 residents. It shows a weak positive association with apprentice counts (r = 0.14), suggesting some influence.

Urbanization (Predictor): Also right-skewed, with no meaningful linear correlation with apprentices (r = –0.02). However, non-linear or regional effects may be present.

Distance and Urbanization: Strongly negatively correlated (r = –0.51), suggesting potential multicollinearity if both variables are included in the model

The histograms help identify skewed distributions, potential outliers (e.g., Midlothian), and the overall data spread—informing model choice and interpretation. The matrix confirms distance as the most informative predictor and highlights overlap with urbanization, warranting caution in joint interpretation.

Bivariate Analysis

To examine the influence of direction and the effect of log-transformations on the predictors, we present the following scatterplots.

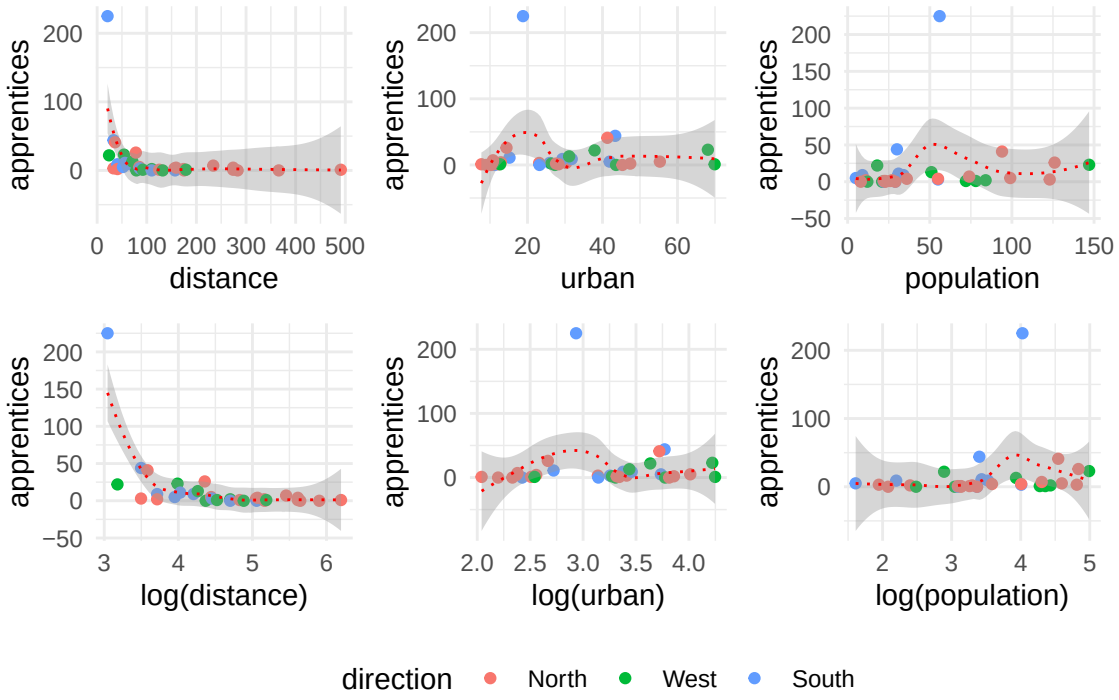

Figure 2: Apprentices by Region and Demographic Factors

The lower row of scatterplots presents the same relationships using log-transformed predictors, which help linearize skewed distributions and stabilize variance. The effect of distance becomes more regular, reinforcing its role as a strong negative predictor. Urbanization still shows no consistent pattern, though subtle regional differences persist. For population, the log transformation slightly clarifies the weak positive trend, but regional variability remains high.

These refined plots illustrate the utility of log transformations in preparing variables for modeling, and further support the inclusion of interaction terms between direction and distance or population.

Poisson Regression Model

Model Definitions

Let Y_i be the number of apprentices in region i, assumed to follow a Poisson distribution:

$$Y_i \sim \text{Poisson}(\lambda_i), \quad \log(\lambda_i) = \eta_i$$

Model fitting is performed via **maximum likelihood estimation**

We build several models to compare the influence of the predictors in the dataset. The linear predictors η_i for the six Poisson models are defined as:

Model 0: $\eta_i = \beta_0 + \beta_1 \cdot d_i + \beta_2 \cdot p_i + \beta_3 \cdot \text{Dir}_i$

Model 1: $\eta_i = \beta_0 + \beta_1 \cdot d_i + \beta_2 \cdot p_i + \beta_3 \cdot \text{Dir}_i + \beta_4 \cdot (p_i \times \text{Dir}_i)$

Model 2: $\eta_i = \beta_0 + \beta_1 \cdot d_i + \beta_2 \cdot \text{Dir}_i + \beta_3 \cdot (d_i \times \text{Dir}_i) + \beta_4 \cdot p_i$

Model 3: $\eta_i = \beta_0 + \beta_1 \cdot \log(d_i) + \beta_2 \cdot \text{Dir}_i + \beta_3 \cdot (\log(d_i) \times \text{Dir}_i) + \beta_4 \cdot \log(p_i) + \beta_5 \cdot u_i$

Model 4: $\eta_i = \beta_0 + \beta_1 \cdot \log(d_i) + \beta_2 \cdot \text{Dir}_i + \beta_3 \cdot (\log(d_i) \times \text{Dir}_i) + \beta_4 \cdot \log(p_i)$

Model 5: $\eta_i = \beta_0 + \beta_1 \cdot \log(d_i) + \beta_2 \cdot \log(p_i)$

Zero-Inflated Poisson Model: $Y_i \sim \begin{cases} 0 & \text{with probability } \pi_i \\ \text{Poisson}(\eta_i) & \text{with probability } 1 - \pi_i \end{cases}$

$$\text{with } \eta_i = \beta_0 + \beta_1 \cdot d_i + \beta_2 \cdot \text{Dir}_i + \beta_3 \cdot (d_i \times \text{Dir}_i) + \beta_4 \cdot p_i$$

and a constant zero-inflation term modeled by a logistic function: $\text{logit}(\pi_i) = \gamma_0$

Abbreviation Key:
`Int` = Intercept, `d` = distance, `p` = population, `dW` = directionWest, `dS` = directionSouth, `p:dW` = population × directionWest, `p:dS` = population × directionSouth, `d:dW` = distance × directionWest, `d:dS` = distance × directionSouth, `log(d)` = log(distance), `log(p)` = log(population), `log(d):dW` = log(distance) × directionWest,`log(d):dS` = log(distance) × directionSouth.

Each model adds complexity in a controlled way to assess trade-offs between interpretability and fit quality (via AIC/log-likelihood). The goal is to identify the most parsimonious model that captures relevant spatial and demographic effects.

- Model 0–2 include linear predictors and interaction terms to capture spatial heterogeneity.
- Model 3–5 introduce logarithmic transformations to reduce skewness and potentially improve fit.
- The **ZIP model** is used to account for **excess zeros**, which may arise in rural or low-population regions with no apprentices at all.

Model Fitting

The model is fitted via **Maximum Likelihood Estimation (MLE)**, maximizing the Poisson likelihood:

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}$$

Table 3: Table: Summary of Model Fit Statistics

Model	AIC	LogLik	df
Model 0	442.7	-216.3	28
Model 1	323.3	-154.7	26
Model 2	212.7	-99.3	26
Model 3	162.2	-73.1	25
Model 4	162.8	-74.4	26
Model 5	259.6	-126.8	30
ZIP Model	214.7	-99.3	NA

Model Assumptions

Poisson regression relies on the following key assumptions (Dobson & Barnett, 2018; Agresti, 2007):

- The response variable is a count (non-negative integers).
- Observations are independent.
- The log of the expected count is a linear function of the predictors.
- The conditional mean equals the conditional variance (equidispersion).

We test these assumptions using diagnostic plots and overdispersion analysis (see below).

Model Assessment

To compare models, we use:

1. Akaike Information Criterion (AIC)

$$\text{AIC} = 2k - 2\log\hat{\mathcal{L}}$$

where: - k is the number of parameters (degrees of freedom), - $\hat{\mathcal{L}}$ is the maximized likelihood.

Lower AIC values indicate better model performance while penalizing complexity.

2. Degrees of Freedom (df)

$$\text{df} = n - k$$

where: - n is the number of observations, - k is the number of estimated parameters.

Table 3 provides a summary of the fitted models. Among all models tested, Model 3—which includes log-transformed distance, direction, urbanization, and their interactions—had the lowest AIC (162.18) and residual deviance, indicating the best overall fit. Simpler models, such as Model 0 (AIC = 442.7), performed significantly worse. Model 2, which includes a distance × direction interaction, showed improved fit (AIC = 212.7) but was still outperformed by the log-transformed models. The Zero-Inflated Poisson (ZIP) model achieved a similar fit to Model 2 (AIC = 214.7) but introduced additional complexity by modeling excess zeros.

Although the dataset contains many zeros, a Vuong test comparing the standard Poisson model to a Zero-Inflated Poisson (ZIP) model found no significant improvement (z = 0.91, p = 0.18). AIC- and BIC-corrected statistics strongly favored the standard model. This suggests that the zeros are well explained by existing covariates, and modeling zero inflation adds unnecessary complexity.

The residual diagnostic plots for the best Models (AIC), which are model 1–3 (shown in Figure X) include:

- Residuals vs Fitted: to assess non-linearity
- Q-Q Plots: to assess normality of residuals (not assumed in Poisson but helps diagnose outliers)
- Scale-Location: to examine homoscedasticity
- Cook's Distance: to identify influential observations

Model 1 shows wider spread in residuals, some curvature in the residual vs fitted plot, and a few high-leverage points (Cook's distance). Model 2 improves upon this, with more linear residual patterns and lower influence points. Model 3 shows the best residual behavior—residuals are tightly clustered, with minimal deviation from linearity and no apparent outliers or influential observations.

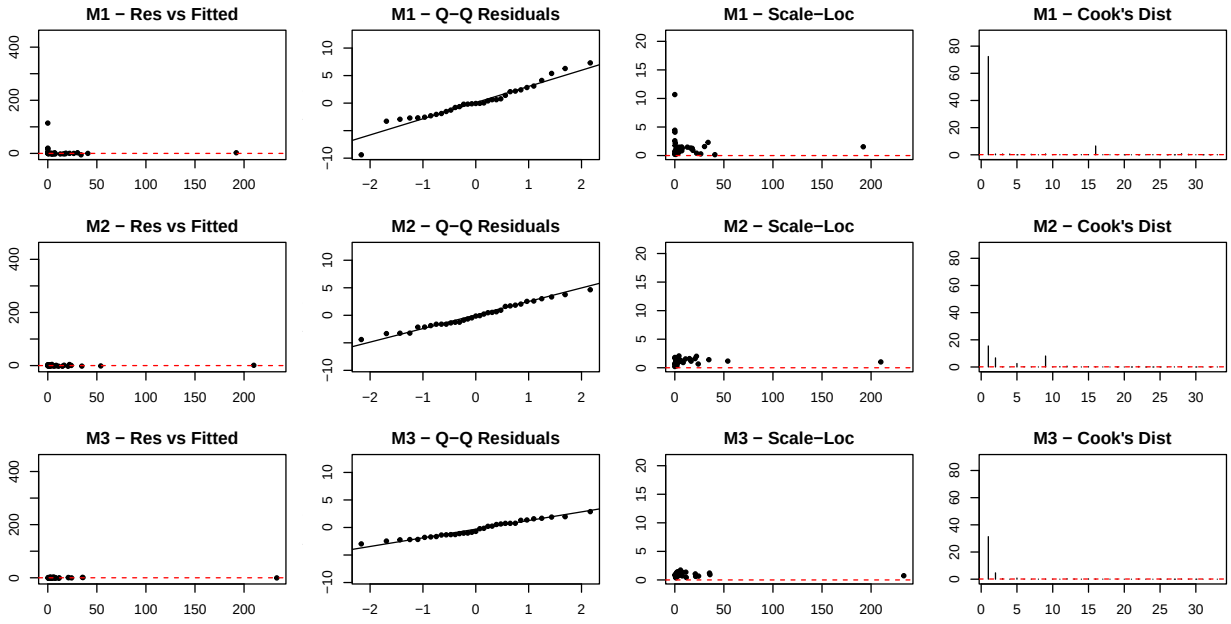

Figure 3: Diagnostic plots for Models M3 to M5

Overdispersion Analysis

A dispersion value close to 1 indicates that the Poisson assumption (mean variance) holds. Values substantially greater than 1 suggest **overdispersion**, meaning the model underestimates variability in the data.

The computed dispersion statistics are:

Dispersion values — Model 0: 7200.29, Model 1: 532.19, Model 2: 4.64, **Model 3: 1.85**, Model 4: 1.93, Model 5: 8.45, ZIP Model: 4.81

Interpretation

- **Model 0** and **Model 1** suffer from extreme overdispersion, confirming they are inadequate for modeling this data.
- **Model 2** shows considerable improvement, but a dispersion of 4.64 still indicates a poor fit.
- **Models 3 and 4** yield dispersion values close to 2, suggesting that these models manage variability relatively well and are more appropriate for inference.
- **Model 5** shows renewed overdispersion (8.45), suggesting that its added complexity may not translate into improved fit.
- The **ZIP model**, despite explicitly modeling excess zeros, still shows overdispersion (4.81) and, did not outperform the standard Poisson model.

Conclusion

Model 3 provides the best balance between goodness of fit and parsimony. It substantially reduces overdispersion while maintaining interpretability and a strong AIC performance. This reinforces the conclusion that the standard Poisson model, with appropriate transformations and interactions, is sufficient and preferable to more complex or zero-inflated alternatives. The dispearsion analyse confirms the choice of poisson regression and dicard the idea of using a different model like a quasi-poisson model or a negative binomial.

Final Model

Final Model Specification

The final Poisson regression model with a log-link function is given by:

$$\begin{aligned}\widehat{y}_i = \exp\Big(&81 + 0.28 \cdot \log(d_i) + 34.76 \cdot \text{dir}_{\text{West},i} \\ &+ 198.90 \cdot \text{dir}_{\text{South},i} + 2.46 \cdot \log(\text{pop}_i) + 0.99 \cdot u_i \\ &+ 0.39 \cdot \log(d_i) \cdot \text{dir}_{\text{West},i} + 0.29 \cdot \log(d_i) \cdot \text{dir}_{\text{South},i}\Big)\end{aligned}$$

Where: - log(d): log-transformed distance from Edinburgh - dirWest , dirSouth : dummy variables for direction (North is baseline) - log(pop)‘: log-transformed regional population - u : urbanization score - Interaction terms account for varying distance effects across directions

- Population: As county population increases, the number of apprentices rises significantly, showing a strong positive effect.
- Urbanization: More urbanized counties send slightly fewer apprentices, suggesting urban areas may offer local alternatives.
- Distance × Direction: Apprentice counts drop sharply with increasing distance, especially in the South. At equal distances, southern counties send more apprentices than western or northern ones.

This highlights population size as the main driver, with distance and geography shaping access

Conclusion

This report explored the geographic and demographic factors influencing apprentice migration to Edinburgh between 1775 and 1799, using Poisson regression modeling. After evaluating several model specifications, Model 3, which includes log-transformed distance and

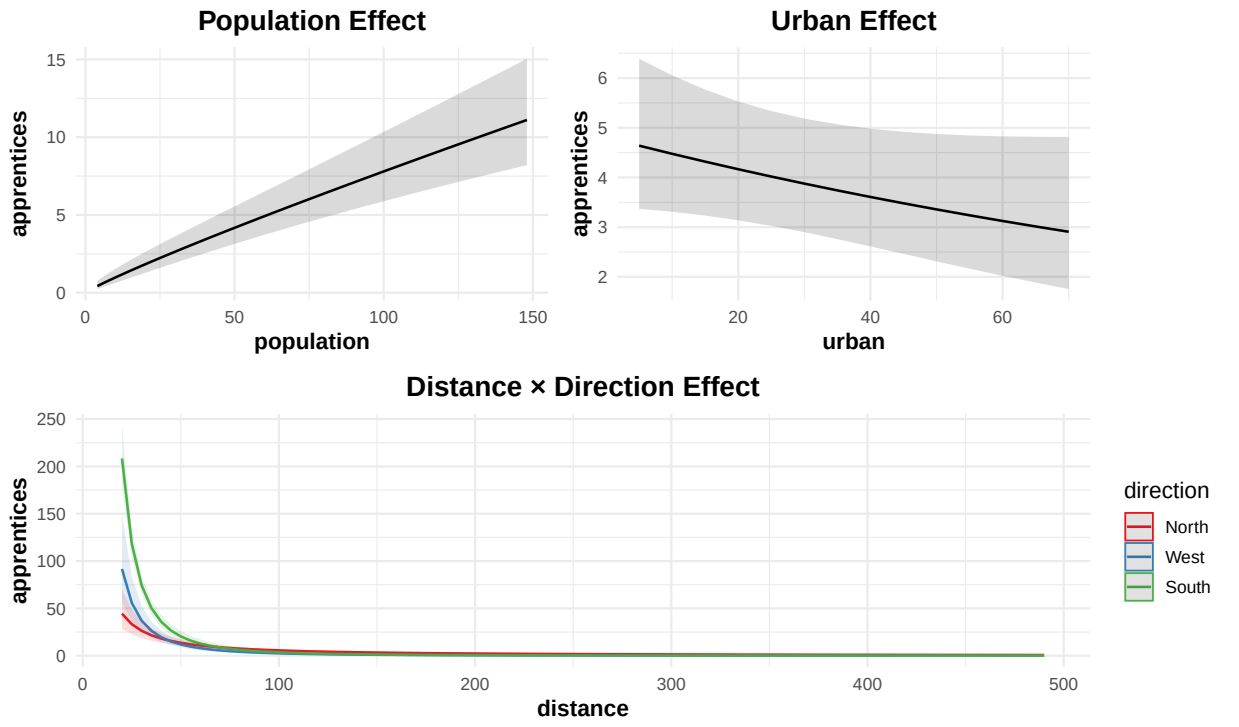

Figure 4: Marginal effects of population, urbanization, and distance on apprentice counts

population, urbanization, and interaction terms with direction, emerged as the best-fitting and most interpretable model.

Key findings include:

- Population size is the strongest positive driver of apprentice counts.
- Urbanization shows a modest negative association, suggesting urban regions may offer local opportunities that reduce outward migration.
- Distance significantly decreases apprentice counts, with the effect varying by region—southern counties send more apprentices than northern or western counties at comparable distances.

Despite the presence of many zeros in the data, zero-inflated models did not provide a better fit, confirming that a well-specified standard Poisson model is sufficient. Diagnostic checks and dispersion analysis further support the robustness of the chosen model.

Overall, this study highlights how spatial accessibility and local demographics jointly shaped labor mobility in historical Scotland. The final model offers both explanatory power and historical insight into the distribution of apprenticeship opportunities during this period. This analysis identifies statistical associations, not causal relationships, between regional characteristics and apprentice counts.

References

- Lovett, A., & Flowerdew, R. (1989). Analysis of Count Data Using Poisson Regression. *The Professional Geographer*, 41(2), 190–198. https://doi.org/10.1111/j.0033-0124.1989.00190.x.
- UCLA Institute for Digital Research and Education. (n.d.). Poisson Regression in R.
- Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. Wiley.
- Dobson, A. J., & Barnett, A. G. (2018). *An Introduction to Generalized Linear Models*. CRC Press.
- Statistics Globe. (2022). Poisson Regression in R (Generalized Linear Model) – YouTube Video.