Poisson Regression Analysis of Apprentice Migration

Nino Gerber

2025-05-07

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 objective is to model the count using a **Poisson regression** and assess the strength and direction of association for each predictor.

Data Description

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)

Data Import and Preprocessing

```
url <- "http://users.stat.ufl.edu/~winner/data/apprentice.dat"
widths <- c(20, 3, -4, 4, -4, 4, -3, 5, -7, 1)
col_names <- c("region", "distance", "apprentices", "population", "urban", "direction")
data <- read.fwf(url, widths = widths, col.names = col_names, strip.white = TRUE)
data$direction <- factor(data$direction, levels = c(1, 2, 3), labels = c("North", "West"
summary(data)
```

```
##     region             distance      apprentices       population
##  Length:33          Min.   : 21.0   Min.   :  0.00   Min.   :  5.00
##  Class :character   1st Qu.: 54.0   1st Qu.:  1.00   1st Qu.: 22.00
##  Mode  :character   Median : 92.0   Median :  3.00   Median : 30.00
##                     Mean   :131.8   Mean   : 14.18   Mean   : 46.58
##                     3rd Qu.:174.0   3rd Qu.:  9.00   3rd Qu.: 72.00
##                     Max.   :491.0   Max.   :225.00   Max.   :147.00
##      urban       direction
##  Min.   : 7.70   North:16
##  1st Qu.:12.90   West : 8
##  Median :27.30   South: 9
##  Mean   :28.62
##  3rd Qu.:41.30
##  Max.   :69.90
```

```
glimpse((data))
```

```
## Rows: 33
## Columns: 6
## $ region      <chr> "Midlothian", "West Lothian", "East Lothian", "Kinross", "~
## $ distance    <int> 21, 24, 33, 33, 36, 41, 41, 52, 54, 56, 67, 71, 78, 79, 85~
## $ apprentices <int> 225, 22, 44, 3, 41, 9, 2, 5, 23, 11, 9, 13, 26, 0, 5, 3, 1~
## $ population  <int> 56, 18, 30, 7, 94, 9, 11, 5, 147, 31, 34, 51, 126, 21, 99,~
## $ urban       <dbl> 18.8, 37.9, 43.4, 30.3, 41.3, 29.3, 47.4, 41.9, 68.1, 15.2~
## $ direction   <fct> South, West, South, North, North, South, North, South, Wes~
```

```
summary_df <- summary(data) %>%
  as.data.frame()

knitr::kable(summary_df, caption = "Summary Statistics of the Apprentice Dataset")
```

Table 1: Summary Statistics of the Apprentice Dataset

Var1	Var2	Freq
	region	Length:33
	region	Class :character
	region	Mode :character
	region	NA
	region	NA
	region	NA
	distance	Min. : 21.0
	distance	1st Qu.: 54.0

Var1	Var2	Freq
	distance	Median : 92.0
	distance	Mean :131.8
	distance	3rd Qu.:174.0
	distance	Max. :491.0
	apprentices	Min. : 0.00
	apprentices	1st Qu.: 1.00
	apprentices	Median : 3.00
	apprentices	Mean : 14.18
	apprentices	3rd Qu.: 9.00
	apprentices	Max. :225.00
	population	Min. : 5.00
	population	1st Qu.: 22.00
	population	Median : 30.00
	population	Mean : 46.58
	population	3rd Qu.: 72.00
	population	Max. :147.00
	urban	Min. : 7.70
	urban	1st Qu.:12.90
	urban	Median :27.30
	urban	Mean :28.62
	urban	3rd Qu.:41.30
	urban	Max. :69.90
	direction	North:16
	direction	West : 8
	direction	South: 9
	direction	NA
	direction	NA
	direction	NA

The dataset consists of 33 Scottish regions, each characterized by five variables. The distance from Edinburgh ranges from 21 to 491 miles, with a mean of 131.8 and a median of 92 miles, suggesting a right-skewed distribution. The population of the regions varies from 5,000 to 147,000, with a mean of 46,580 and a median of 30,000, indicating a few highly populated regions. The urbanization score spans 7.7 to 69.9, with an average of 28.6, and shows a spread across rural to semi-urban areas. The direction from Edinburgh is a categorical variable with 16 regions to the North, 8 to the West, and 9 to the South. The response variable, apprentices, has a minimum of 0 and a maximum of 225, with a mean of 14.2 and a median of 3, indicating a highly skewed distribution dominated by a few large values — especially one outlier (Midlothian) sending over 200 apprentices.

Exploratory Data Analysis

Univariate Analysis

```
p1 <- ggplot(data, aes(x = apprentices)) +
  geom_histogram(binwidth = 10, fill = "lightblue", color = "black") +
  labs(title = "Distribution of Apprentices", x = "Number of Apprentices")

p2 <- ggplot(data, aes(x = distance)) +
  geom_histogram(binwidth = 20, fill = "lightgreen", color = "black") +
  labs(title = "Distribution of Distance", x = "Distance (miles)")

p3 <- ggplot(data, aes(x = population)) +
  geom_histogram(binwidth = 10, fill = "lightcoral", color = "black") +
  labs(title = "Distribution of Population", x = "Population (in thousands)")

p4 <- ggplot(data, aes(x = urban)) +
  geom_histogram(binwidth = 5, fill = "lightgoldenrod", color = "black") +
  labs(title = "Distribution of Urbanization Score", x = "Urbanization")

(p1 | p2) /
(p3 | p4)
```

Univariate Distributions

To understand the structure of our dataset before modeling, we examined the distribution of each variable using histograms. These provide insight into the shape, spread, and potential issues in the data.

- **Apprentices (Response Variable):**
  The distribution is highly right-skewed, with most counties sending very few apprentices. A single outlier (likely Midlothian) sent over 200 apprentices. This justifies the use of a Poisson regression model, which is designed for skewed count data.

- **Distance (Predictor):**
  Many counties are located within 50–150 miles of Edinburgh, with fewer in the extreme distances (200–500 miles). This shows that the data is concentrated in a realistic and informative range for modeling accessibility effects.

- **Population (Predictor):**
  The population distribution is also right-skewed, with most counties having fewer than 50,000 people. This informs us that a few highly populated counties may dominate the influence on apprentice numbers. Since the Poisson model uses a log link, it can handle this skew naturally, but we must remain cautious of leverage effects.

- **Urbanization (Predictor):**
  Urbanization scores show a somewhat bimodal distribution, with groups of counties

either mostly rural or moderately urban. This suggests urbanization might not have a simple linear relationship with apprentice migration and may interact with other factors such as geographic direction.

These histograms are crucial for identifying skewed distributions, potential outliers, and understanding where most of the data lies. This informs both our choice of model and interpretation of its output.

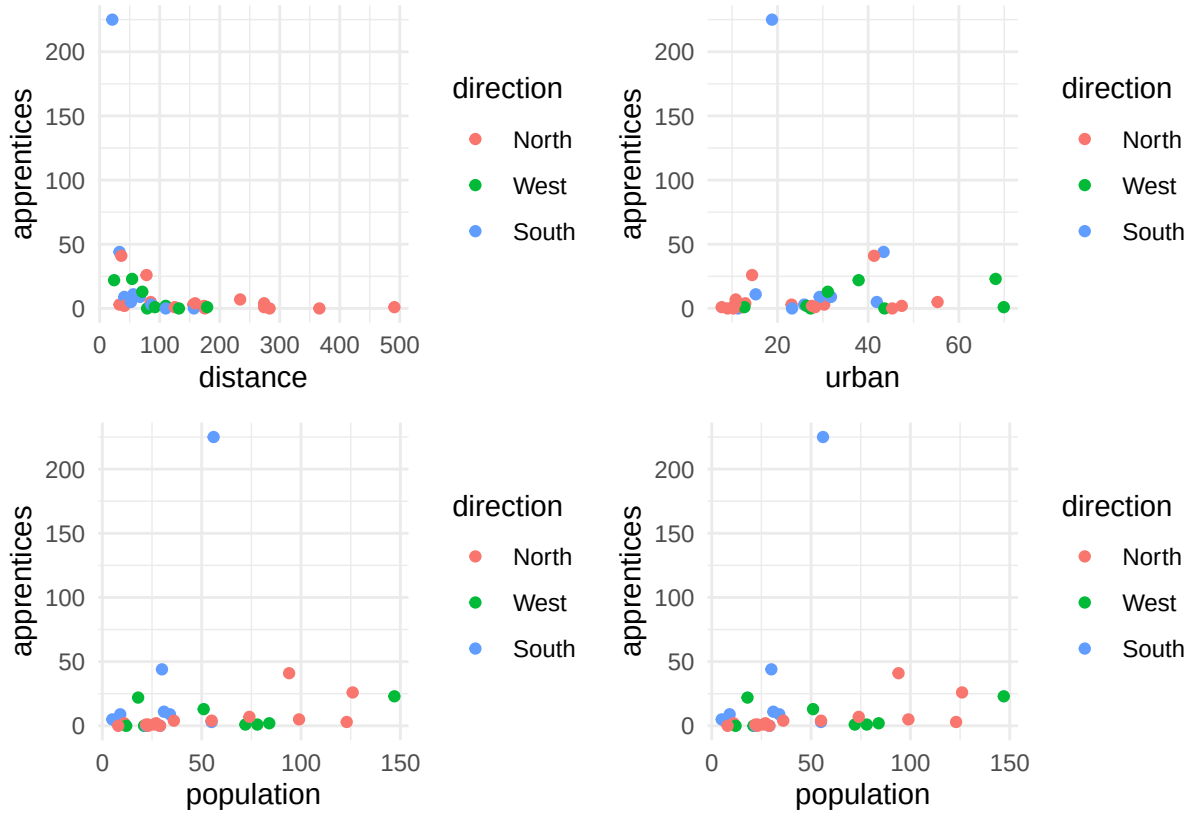
```
pl1 <- ggplot(data, aes(y = apprentices, x = distance, color = direction)) +
  geom_point(alpha = 1) +
  theme_minimal()

pl2 <- ggplot(data, aes(y = apprentices, x = urban, color = direction)) +
  geom_point(alpha = 1) +
  theme_minimal()

pl3 <- ggplot(data, aes(y = apprentices, x = population, color = direction)) +
  geom_point(alpha = 1) +
  theme_minimal()

pl4 <- ggplot(data, aes(y = apprentices, x = population, color = direction)) +
  geom_point(alpha = 1) +
  theme_minimal()

(pl1 | pl2) /
(pl3 | pl4)
```



Bivariate Relationships by Region

The following scatterplots show the relationships between the number of apprentices and each predictor (distance, urbanization, and population), with points colored by region (`direction`: North, South, West).

- **Distance vs Apprentices:**
  There is a clear negative relationship: counties farther from Edinburgh tend to send fewer apprentices. This is especially visible among Northern counties (red), which are spread across larger distances and mostly have low apprentice counts. In contrast, many Southern counties (blue) are closer and show higher apprentice numbers, indicating that `direction` modifies the effect of distance. This supports the inclusion of both `distance` and `direction` in the model and suggests a possible interaction.

- **Urbanization vs Apprentices:**
  No strong linear trend is visible across regions. Apprentice counts remain low or modest regardless of the degree of urbanization. One major outlier from the South stands out with high apprentice numbers and low urbanization. This suggests urbanization may not be a strong independent predictor and could be confounded or interacting with other variables such as population or direction.

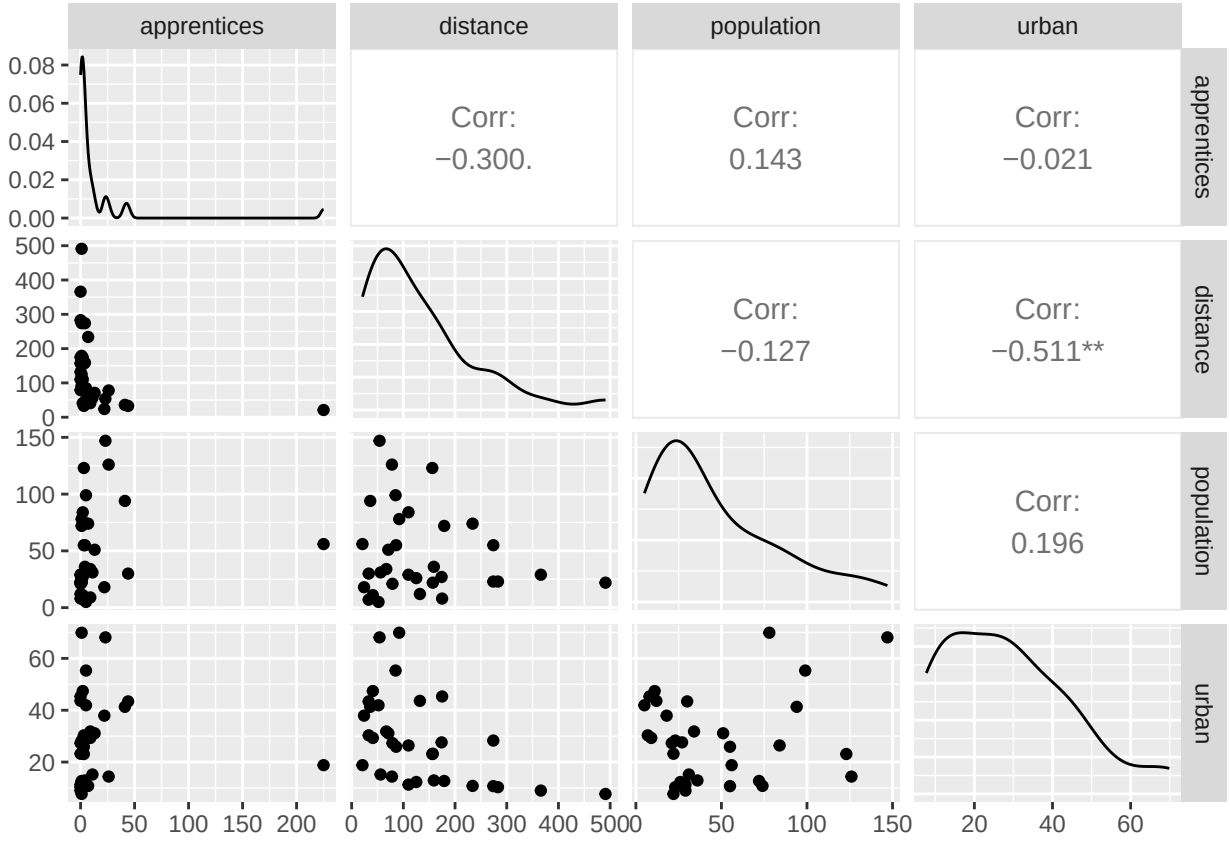
- **Population vs Apprentices:**
  A weak positive association is observed: counties with larger populations tend to send

slightly more apprentices. However, the relationship is not consistent across regions, and large variability is present. The same Southern outlier dominates the upper end of apprentice counts. This supports including `population` in the model but cautions us to monitor its influence.

These plots highlight how `direction` affects relationships between predictors and the response. In particular, Southern counties are visibly different in their behavior, supporting the inclusion of `direction` as a categorical variable and motivating interaction terms in the Poisson model.

Bivariate Analysis

```
ggpairs(data[, c("apprentices", "distance", "population", "urban")])
```



Correlation Matrix and Bivariate Density Plots

The correlation matrix below provides a compact overview of the linear relationships between all numerical variables in the dataset. It includes:

- **Histograms/density plots on the diagonal** (univariate distributions),
- **Scatterplots with smoothing curves** in the lower triangle (bivariate relationships),
- **Pearson correlation coefficients** in the upper triangle.

Key Observations:

- **Apprentices vs Distance**:
  The correlation is –0.300, indicating a moderate negative relationship. The scatterplot confirms that counties farther from Edinburgh tend to send fewer apprentices, justifying the inclusion of `distance` as a strong predictor.
- **Apprentices vs Population**:
  The correlation is 0.143, suggesting a weak positive association. While more populous counties may tend to send more apprentices, the relationship is not strong and shows substantial variation.
- **Apprentices vs Urbanization**:
  The correlation is near zero (–0.021), indicating no meaningful linear association. However, some nonlinearity might exist or effects could vary by region, which may explain why `urban` was still significant in the Poisson model.
- **Distance vs Urbanization**:
  A strong negative correlation (–0.511**) is observed, implying that counties closer to Edinburgh tend to be more urbanized. This introduces potential multicollinearity when both variables are included in a model and should be taken into account when interpreting coefficients.
- **Other correlations** (e.g., distance vs population, population vs urban) are relatively weak, suggesting that predictors are not strongly collinear beyond the noted pair.

Conclusion:

This matrix is valuable for identifying which variables are likely to be informative in the model and which ones may overlap. It confirms `distance` as a key predictor and flags the strong inverse relationship between `distance` and `urbanization`, guiding caution in model interpretation.

Poisson Regression Model

Model Definition

Let Y_i be the number of apprentices from region i. We assume:

$$Y_i \sim \text{Poisson}(\lambda_i), \quad \log(\lambda_i) = \beta_0 + \beta_1 \cdot \text{Distance}_i + \beta_2 \cdot \text{Population}_i + \beta_3 \cdot \text{Urban}_i + \beta_4 \cdot \text{Direction}_i$$

Model fitting is performed using **maximum likelihood estimation** via `glm()` with `family = poisson(link = "log")`.

Model Fitting

```
model0 <- glm(apprentices ~ distance + population + direction, data = data, family = poi
model1 <- glm(apprentices ~ distance + population*direction , data = data, family = pois
model2 <- glm(apprentices ~ distance * direction + population , data = data, family = po
summary(model0)
```

```
##
## Call:
## glm(formula = apprentices ~ distance + population + direction,
##     family = poisson(link = "log"), data = data)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     2.963893   0.210050  14.110  < 2e-16 ***
## distance       -0.031851   0.002022 -15.753  < 2e-16 ***
## population      0.021593   0.001627  13.271  < 2e-16 ***
## directionWest  -0.614378   0.168156  -3.654 0.000259 ***
## directionSouth  1.429917   0.145045   9.858  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1350.44  on 32  degrees of freedom
## Residual deviance:  338.33  on 28  degrees of freedom
## AIC: 442.68
##
## Number of Fisher Scoring iterations: 7
```

```
summary(model1)
```

```
##
## Call:
## glm(formula = apprentices ~ distance + population * direction,
##     family = poisson(link = "log"), data = data)
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                2.735607   0.267057  10.244  < 2e-16 ***
## distance                  -0.025773   0.001812 -14.223  < 2e-16 ***
## population                 0.020234   0.002703   7.487 7.06e-14 ***
## directionWest              0.661409   0.338249   1.955 0.050537 .
```

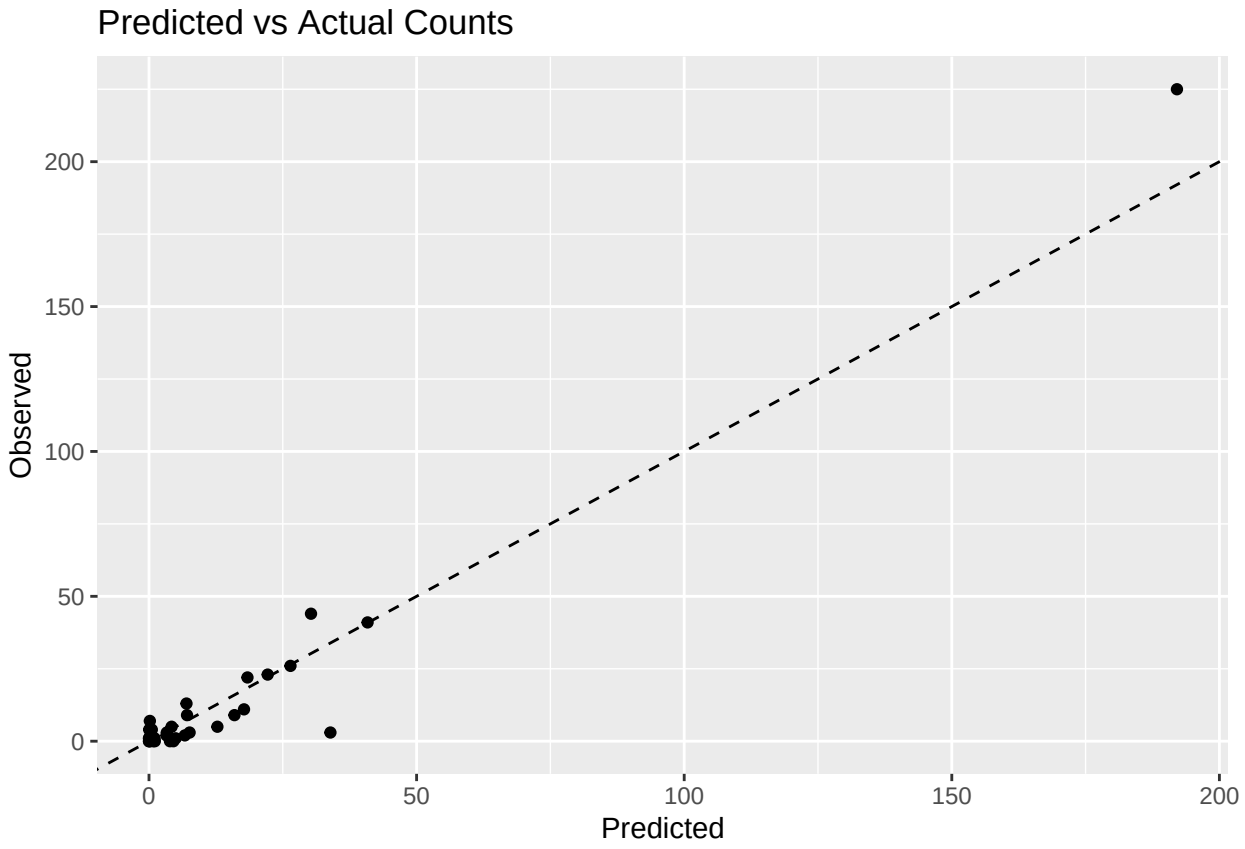
```
## directionSouth            -0.249929   0.345510  -0.723 0.469456
## population:directionWest  -0.012793   0.003499  -3.656 0.000256 ***
## population:directionSouth  0.038934   0.005389   7.224 5.03e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1350.44  on 32  degrees of freedom
## Residual deviance:  214.98  on 26  degrees of freedom
## AIC: 323.33
##
## Number of Fisher Scoring iterations: 6
```

```
summary(model2)
```

```
##
## Call:
## glm(formula = apprentices ~ distance * direction + population,
##     family = poisson(link = "log"), data = data)
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)               2.031386   0.229790   8.840  < 2e-16 ***
## distance                 -0.009483   0.001605  -5.907 3.48e-09 ***
## directionWest             1.460426   0.383228   3.811 0.000138 ***
## directionSouth            4.176493   0.236034  17.694  < 2e-16 ***
## population                0.014933   0.001596   9.354  < 2e-16 ***
## distance:directionWest   -0.030146   0.006399  -4.711 2.47e-06 ***
## distance:directionSouth -0.071339   0.005772 -12.359  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1350.44  on 32  degrees of freedom
## Residual deviance:  104.34  on 26  degrees of freedom
## AIC: 212.68
##
## Number of Fisher Scoring iterations: 5
```

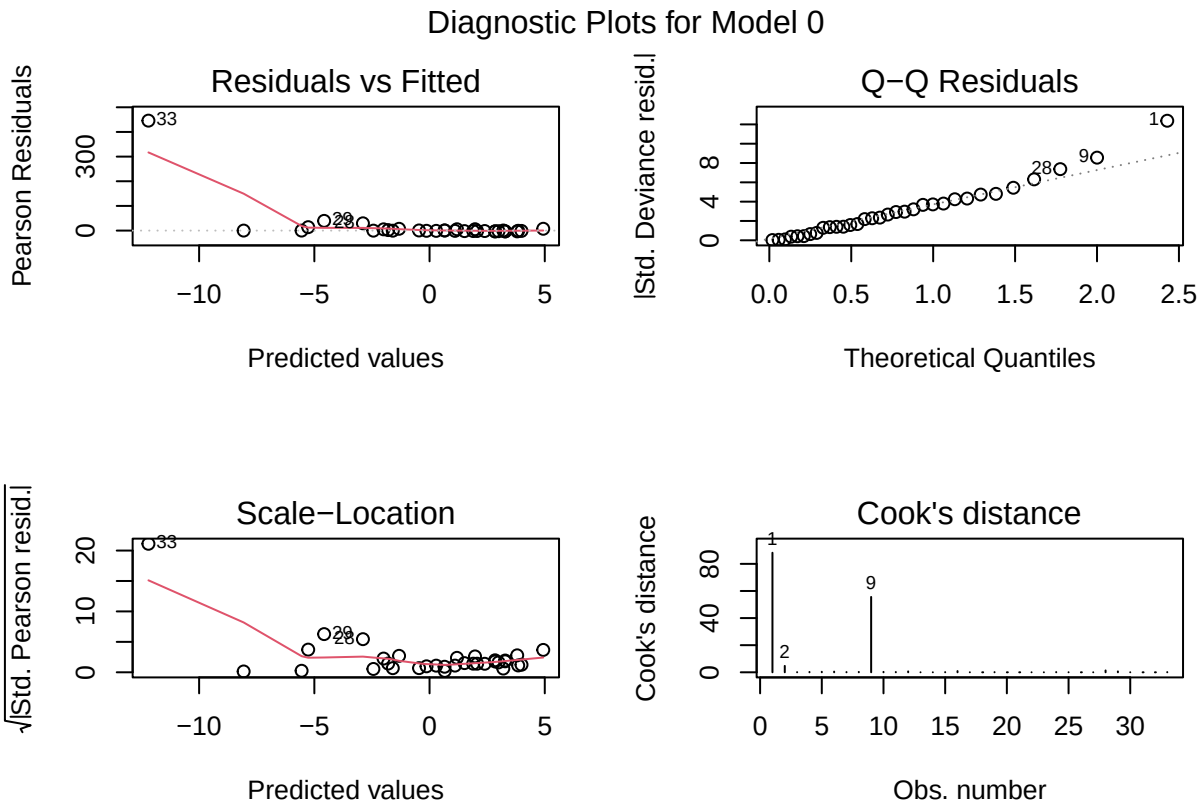
Model Assessment

```
data$pred <- predict(model1, type = "response")
ggplot(data, aes(x = pred, y = apprentices)) +
  geom_point() +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed") +
  labs(title = "Predicted vs Actual Counts",
       x = "Predicted", y = "Observed")
```

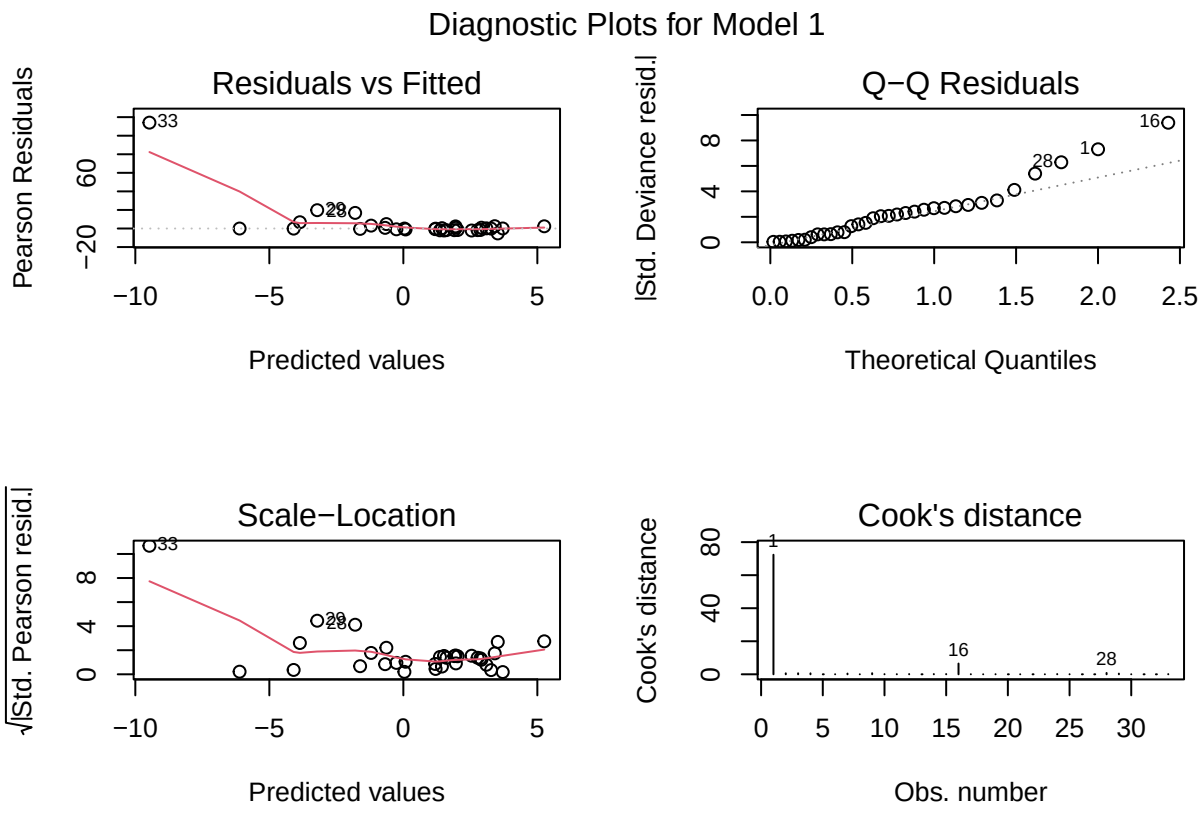


Assumptions and Diagnostics

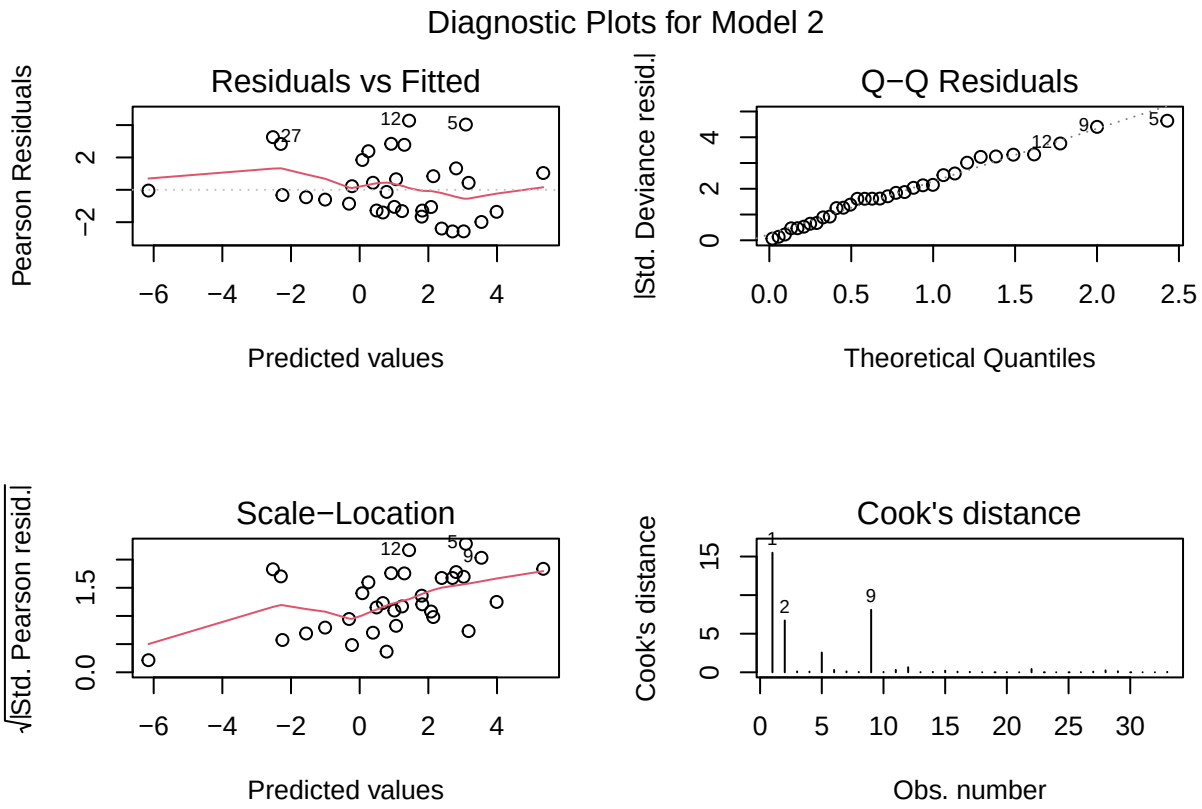
```
par(mfrow = c(2, 2))
# Plot for model0
plot(model0, which = 1:4)
mtext("Diagnostic Plots for Model 0", side = 3, line = -2, outer = TRUE)
```




```r
# Plot for model1
plot(model1, which = 1:4)
mtext("Diagnostic Plots for Model 1", side = 3, line = -2, outer = TRUE)
```




```r
# Plot for model2
plot(model2, which = 1:4)
mtext("Diagnostic Plots for Model 2", side = 3, line = -2, outer = TRUE)
```



Diagnostic Evaluation of Poisson Models

Diagnostic plots reveal substantial improvements in model fit across model versions. Model 0 and Model 1 exhibit severe residual issues, with observation 33 (likely Midlothian) emerging as an extreme outlier and highly influential point. The residuals vs fitted plots show strong deviation from expected patterns, and the Q-Q plots reveal heavy upper tails, indicating poor residual normality.

In contrast, Model 2 shows a significant improvement: residuals are more centered and homoscedastic, the influence of outliers is reduced, and Cook's distance values are substantially lower. This suggests Model 2 handles outlier influence better and achieves more stable and interpretable parameter estimates.

Overall, Model 2 is favored not only in terms of statistical fit (lower AIC and deviance) but also based on diagnostic validity.

The dispersion statistic is calculated to assess whether the variance significantly exceeds the mean, which would suggest overdispersion and the need for an alternative model.

Final Model

```r
# Calculate dispersion for model0
dispersion0 <- sum(residuals(model0, type = "pearson")^2) / model0$df.residual

# Calculate dispersion for model1
dispersion1 <- sum(residuals(model1, type = "pearson")^2) / model1$df.residual

# Calculate dispersion for model2
dispersion2 <- sum(residuals(model2, type = "pearson")^2) / model2$df.residual

# Print results
dispersion0
```

```
## [1] 7200.291
```

```r
dispersion1
```

```
## [1] 532.1909
```

```r
dispersion2
```

```
## [1] 4.638105
```

Overdispersion Analysis

To assess whether the Poisson regression models are appropriate, we calculated the dispersion statistic for each model using the Pearson residuals:

$$\text{Dispersion} = \frac{\sum (\text{Pearson residuals})^2}{\text{Residual degrees of freedom}}$$

The results are as follows:

- **Model 0:** 7200.29
- **Model 1:** 532.19
- **Model 2:** 4.64

A dispersion value significantly larger than 1 indicates **overdispersion**, meaning that the variance of the response variable is much greater than the mean—violating the assumption of the Poisson model.

Interpretation

- **Model 0 and Model 1** show extreme overdispersion, suggesting that the model assumptions are strongly violated and that inference based on standard errors and p-values is unreliable.
- **Model 2** performs significantly better, but the dispersion value of 4.64 still indicates **moderate overdispersion**, suggesting that the model still underestimates variability in the data.

Implications and Next Steps

To correct for overdispersion, we recommend using: - A **quasi-Poisson** model, which retains the structure of the model but adjusts standard errors to be more robust. - Or preferably, a **Negative Binomial** model, which introduces an additional parameter to explicitly account for extra variability and typically results in more reliable inference.

These adjustments ensure more accurate confidence intervals and hypothesis tests, without substantially changing the estimated coefficients.

```
model2_quasi <- glm(apprentices ~ distance * direction + population,
                    data = data, family = quasipoisson(link = "log"))
summary(model2_quasi)
```

```
##
## Call:
## glm(formula = apprentices ~ distance * direction + population,
##     family = quasipoisson(link = "log"), data = data)
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               2.031386   0.494898   4.105 0.000356 ***
## distance                 -0.009483   0.003458  -2.743 0.010883 *
## directionWest             1.460426   0.825356   1.769 0.088548 .
## directionSouth            4.176493   0.508344   8.216 1.07e-08 ***
## population                0.014933   0.003438   4.343 0.000190 ***
## distance:directionWest   -0.030146   0.013782  -2.187 0.037905 *
## distance:directionSouth -0.071339   0.012431  -5.739 4.84e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 4.638395)
##
##     Null deviance: 1350.44  on 32  degrees of freedom
## Residual deviance:  104.34  on 26  degrees of freedom
```

```
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

```
model2_nb <- glm.nb(apprentices ~ distance * direction + population, data = data)
summary(model2_nb)
```

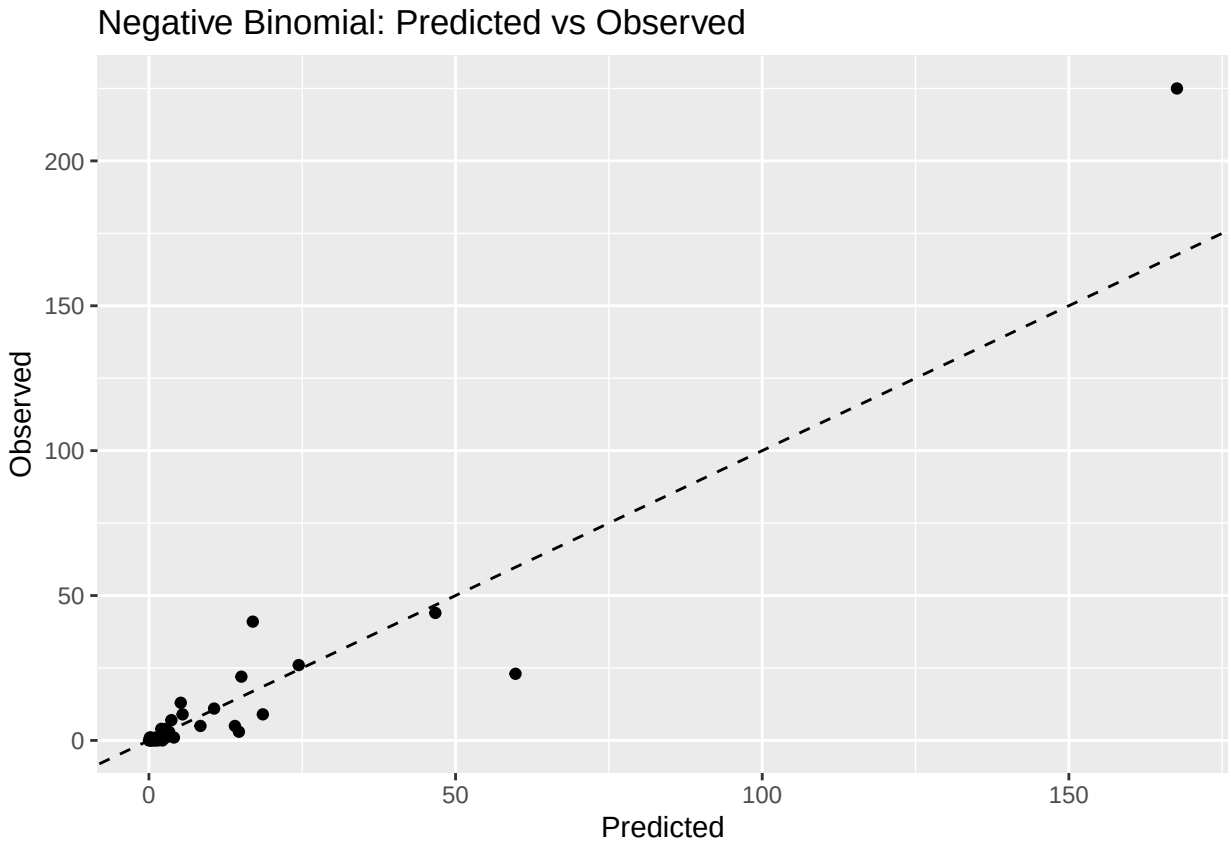
```
##
## Call:
## glm.nb(formula = apprentices ~ distance * direction + population,
##     data = data, init.theta = 3.081508181, link = log)
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                1.246623   0.462450   2.696  0.00702 **
## distance                  -0.005800   0.002188  -2.651  0.00802 **
## directionWest              1.988877   0.776862   2.560  0.01046 *
## directionSouth             4.176705   0.734226   5.689 1.28e-08 ***
## population                 0.019061   0.003571   5.337 9.45e-08 ***
## distance:directionWest    -0.030253   0.009448  -3.202  0.00136 **
## distance:directionSouth -0.059390   0.012123  -4.899 9.63e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(3.0815) family taken to be 1)
##
##     Null deviance: 219.755  on 32  degrees of freedom
## Residual deviance:  33.314  on 26  degrees of freedom
## AIC: 176.29
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  3.08
##           Std. Err.:  1.24
##
##  2 x log-likelihood:  -160.292
```

```
# AIC comparison
AIC(model0, model1, model2, model2_nb)
```

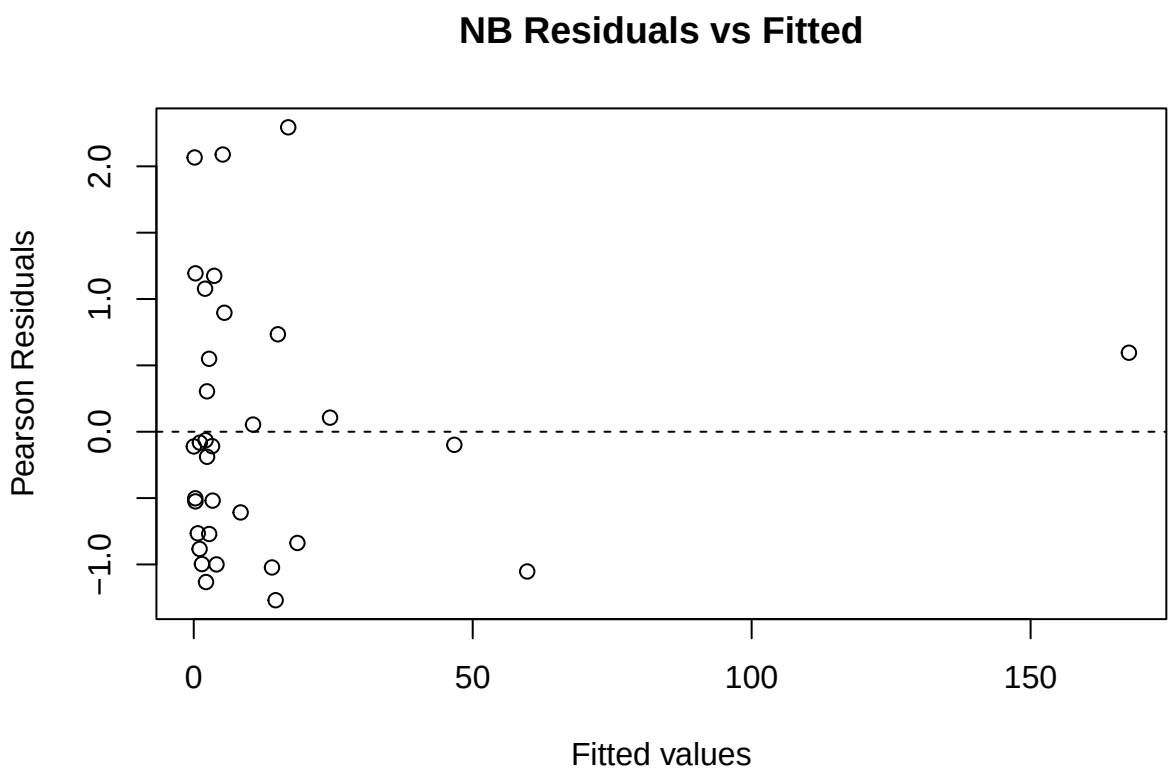
```
##           df      AIC
## model0     5 442.6777
```

```
## model1     7 323.3266
## model2     7 212.6825
## model2_nb  8 176.2917
```

```
data$fit_nb <- predict(model2_nb, type = "response")
ggplot(data, aes(x = fit_nb, y = apprentices)) +
  geom_point() +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed") +
  labs(title = "Negative Binomial: Predicted vs Observed",
       x = "Predicted", y = "Observed")
```




```
resid_nb <- residuals(model2_nb, type = "pearson")
plot(data$fit_nb, resid_nb,
     xlab = "Fitted values", ylab = "Pearson Residuals",
     main = "NB Residuals vs Fitted")
abline(h = 0, lty = 2)
```



The estimated log-linear model is:

$$\widehat{\log(\lambda)} = \text{intercept} + \beta_1 \cdot \text{Distance} + \beta_2 \cdot \text{Population} + \beta_3 \cdot \text{Urbanization} + \beta_4 \cdot \text{Direction (South/West)}$$

Each coefficient represents the **log change** in the expected number of apprentices per unit increase in the predictor.

Conclusions

- **Distance**: Negative and significant → counties farther from Edinburgh tend to send fewer apprentices.
- **Population**: Positive → larger populations send more apprentices.
- **Urbanization**: Positive and significant → more urbanized regions tend to send more apprentices.
- **Direction**: Some directions show significant differences relative to the baseline.

The model identifies meaningful **associations** between regional characteristics and apprentice migration. No causality can be inferred. Dispersion is close to 1, suggesting the Poisson model is appropriate.

References

- 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.