ST 623 – Generalized Regression Models (ST_623_001_F2024)

Modeling Female Sex Worker Distribution and Determinants in Sub-Saharan Africa to Enhance HIV-Prevention Strategies

Oluwaseun Daniel Fowotade

Instructor: Dr. Duo Jiang

Department of Statistics
Oregon State University

December 19, 2024

Introduction

Understanding the factors that influence the distribution of Female Sex Workers (FSWs) is critical for improving HIV prevention strategies in sub-Saharan Africa. This study examines data from five countries—Mozambique, Malawi, Zimbabwe, Zambia, and Botswana—to explore how environmental, sociological, and collection-related factors affect FSW prevalence.

The goal is to determine whether environmental factors, such as urbanization, climate, and access to clean water, or sociological factors, such as crime rates, wealth, and HIV fear, better predict the presence of FSWs. With limited resources, UNAIDS plans to collect only one set of variables—environmental or sociological—in future studies. This research will help guide that decision, ensuring resources are allocated efficiently and interventions are targeted effectively.

By shedding light on these patterns, the study aims to enhance the implementation of HIV prevention services, contributing to more effective public health strategies in the region.

Digging into the World of Female Sex Workers in Sub-Saharan Africa

We analyzed a dataset containing information on Female Sex Workers (FSWs) across five sub-Saharan African countries. The original dataset comprised 750 observations and 20 variables, but 75 entries (10%) had missing information. After cleaning the data, we retained 675 observations (90% of the dataset) for our analysis, ensuring a robust and accurate foundation for statistical exploration.

Our objective was to identify factors influencing the distribution and prevalence of FSWs in different regions. To achieve this, we categorized the explanatory variables into three main groups:

- **Environmental Variables:** This category included factors such as annual rainfall, average temperature, nighttime light activity (`nightlight`), and access to clean water.
- **Sociological Variables:** These captured social and demographic aspects potentially influencing FSW prevalence, including high crime rates (`highCrime`), access to insect nets (`insectNet`), and the average age of first sexual experience (`ageFirstSex`).
- **Collection-Related Variables:** These variables detailed the context of data collection, such as the specific region (`region`), year (`dataYear`), month (`month`), and country (`country`).

This categorization allowed us to explore the interplay between environmental, sociological, and geographic factors and their potential contribution to observed patterns in FSW distribution across the region.

Early Findings

- **FSW Counts:** Counts ranged from 0 to 134, with an average of approximately 5.56 FSWs per observation.
- **Distribution:** Visualizing the distribution of FSW counts (e.g., using histograms) provided insights into the data spread and variability.
- **Correlation Analysis:** Weak negative correlations were observed between FSW counts and variables such as built-up areas (`built`), longer growing seasons (`growingSeason`), higher rainfall (`rain`), nighttime light activity (`nightlight`), and access to clean water (`cleanWater`). Conversely, temperature (`temperature`) showed a positive correlation with FSW counts.

These initial findings suggest that more developed areas with better infrastructure and readily available clean water may be associated with lower FSW prevalence. In contrast, warmer temperatures appear linked to higher FSW counts. However, it is crucial to note that correlation does not imply causation. Further statistical modeling is needed to explore these potential relationships in greater depth.

Visualization and Trend Analysis

To deepen our understanding, we employed visualizations such as scatterplots and correlation matrices to examine relationships between FSW counts, environmental and sociological factors, and geographic locations. These visual tools highlighted potential trends, distributions, and outliers requiring further investigation. They also provided insights into how explanatory variables interact with FSW counts and each other.

Distribution Selection Using Environmental Variables

To explore *FSWCount* as the response variable, an initial Poisson mixed-effects model was developed. This model accounted for spatial and temporal variability by including *country*, *region*, *data year*, and *month* as random effects, alongside environmental variables as fixed covariates. Diagnostic evaluations (Figure 4) highlighted overdispersion issues, as evidenced by the Residuals vs. Fitted plot, which showed increasing variability in residuals with higher fitted values. Additionally, the QQ plot revealed deviations from normality, with residuals displaying heavy-tailed behavior.

An overdispersion test confirmed these findings, reporting a dispersion parameter of $\theta = 6.98$, which exceeded what would be expected under a Poisson distribution. To address this, alternative models were evaluated, including the negative binomial, zero-inflated Poisson, zero-inflated negative binomial, and hurdle models. Model selection was guided by the Akaike Information Criterion (AIC), which balances model fit and complexity. The negative binomial mixed-effects model achieved the lowest AIC (2873.77), indicating it provided the best fit for the data.

Given that approximately 50% of the observations in *FSWCount* were zeros, a zero-inflated negative binomial model was also considered. However, the proportion of excess zeros was negligible, indicating no substantial improvement from using a zero-inflated structure. Thus, the negative binomial mixed-effects model was chosen as the final model.

The model is expressed as follows:

$$\text{FSWCount}_i \sim \text{Negative Binomial}(\mu_i, \eta)$$

$$\log(\mu_i) = \beta_0 + \beta_1 \cdot \text{popDensity}_i + \beta_2 \cdot \text{built}_i + \beta_3 \cdot \text{temperature}_i + \alpha_{\text{country}[i]} + \beta_{\text{region}[i]} + \log(\text{surveyArea}_i)$$

$$\alpha_{\text{country}} \sim N(0, \sigma_\alpha^2),\ \beta_{\text{region}} \sim N(0, \sigma_\beta^2),\ \eta\,(\text{Overdispersion Parameter}).$$

Distribution Selection Using Sociological Variables

A similar approach was applied to analyze *FSWCount* in relation to sociological variables. A Poisson mixed-effects model was initially fitted, incorporating *country*, *region*, *data year*, and *month* as random effects, with sociological variables as fixed covariates. Diagnostic evaluations (Figure 5) again highlighted overdispersion, as seen in the Residuals vs. Fitted plot, where residual variability increased with fitted values. The QQ plot further confirmed deviations from normality, showing heavy-tailed residuals.

To address these issues, alternative models were considered, including negative binomial, zero-inflated Poisson, zero-inflated negative binomial, and hurdle models. Among these, the negative binomial mixed-effects model achieved the lowest AIC (2872.94), making it the best fit. A zero-inflated negative binomial model was also evaluated, but the negligible proportion of excess zeros indicated no meaningful improvement. As a result, the negative binomial mixed-effects model was selected as the final model.

The model is defined as:

$$\text{FSWCount}_i \sim \text{Negative Binomial}(\mu_i, \eta)$$

$$\log(\mu_i) = \beta_0 + \beta_1 \cdot \text{highCrime}_i + \beta_2 \cdot \text{insectNet}_i + \alpha_{\text{country}[i]} + \beta_{\text{region}[i]} + \gamma_{\text{dataYear}[i]} + \log(\text{surveyArea}_i)$$

$$\alpha_{\text{country}} \sim N(0, \sigma_\alpha^2),\ \beta_{\text{region}} \sim N(0, \sigma_\beta^2),\ \gamma_{\text{dataYear}} \sim N(0, \sigma_\gamma^2),\ \eta\,(\text{Overdispersion Parameter}).$$

Variable Selection Using Environmental Variables

The process of selecting variables for the environmental model began with a preliminary model that included all environmental covariates. An ANOVA test was conducted to assess the necessity of random effects, which provided strong evidence for their inclusion ($p < 0.001$). However, random effects for *month* and *data year* were excluded in subsequent iterations due to weak statistical support ($p > 0.2$). The removal of these components resulted in a slight improvement in the AIC score, indicating better model fit.

To refine the model further, residuals were analyzed for systematic patterns. No significant patterns were observed, confirming that random slopes were unnecessary for the included variables. Using Maximum Likelihood (ML) estimation, three covariates—*population density*, *built environment*, and *temperature*—were identified as significant contributors. The inclusion of these variables led to an improved AIC score of 2869.42, enhancing the model's explanatory power.

Variable Selection Using Sociological Variables

For the sociological model, the selection process followed a similar approach. Random effects for *country*, *region*, and *data year* were retained based on strong statistical support, while *month* was excluded due to weak evidence ($p > 0.7$).

Fixed-effect selection identified *high crime* and *insect net usage* as significant predictors of *FSWCount*. These variables demonstrated meaningful relationships with the response variable and were retained in the final model. The optimized model achieved an improved AIC score of 2863.21, confirming its robustness and reliability in capturing the effects of sociological factors.

Results and Conclusion

The analysis successfully identified significant associations between environmental and sociological variables and the distribution of Female Sex Workers (FSWs) across the sampled regions. By employing statistical modeling, we achieved our goal of understanding the factors influencing FSW counts while accounting for both fixed and random effects. This comprehensive approach not only provided robust insights into these relationships but also highlighted key patterns relevant to public health and policy.

Environmental Variables

Among the environmental variables, *population density* (*popDensity*) emerged as a significant predictor of FSW counts, with sparsely populated regions exhibiting higher prevalence. This finding suggests that socio-economic or demographic factors unique to less densely populated areas might influence FSW distribution. Similarly, the *built environment index* showed a negative correlation with FSW counts, indicating that less urbanized areas tend to have higher FSW prevalence. This could reflect disparities in infrastructure, economic development, or access to resources. Moreover, *temperature* demonstrated a positive association with FSW counts, which might indicate region-specific environmental or cultural dynamics affecting FSW activity.

Sociological Variables

Sociological variables also played a critical role in shaping FSW distribution. *High crime rates* (*highCrime*) were significantly associated with increased FSW counts, underscoring the impact of socio-economic vulnerabilities and unsafe environments. Interestingly, *insect net usage* (*insectNet*) was positively correlated with FSW counts. Although the mechanism underlying this relationship is unclear, it may serve as a proxy for regional variations in public health resources, living conditions, or other unmeasured factors.

Random Effects

The random effects structure of the model accounted for unmeasured variability across spatial and temporal dimensions. Random intercepts for *country*, *region*, and *data year* were statistically significant, capturing hierarchical clustering in the data. This reinforces the importance of considering contextual and structural factors when analyzing complex phenomena like FSW distribution.

Model Performance

The final negative binomial mixed-effects model provided the best fit, achieving the lowest AIC score (2863.208). Likelihood ratio tests confirmed the significance of the fixed effects and the necessity of including random effects. This highlights the value of integrating environmental and sociological dimensions to comprehensively understand FSW distribution.

Implications and Future Directions

This study underscores the complex interplay between environmental and sociological variables in shaping FSW distribution. However, these findings are generalizable only to the sampled regions and timeframes or to other contexts with similar characteristics. Given the observational nature of the study, causal inferences cannot be drawn. Nevertheless, the results offer actionable insights for public health interventions and policies, particularly in regions with similar socio-economic and environmental contexts.

Future research should consider incorporating additional covariates, such as economic indicators and access to healthcare, to capture a broader range of influencing factors. Additionally, longitudinal studies could help identify dynamic changes in FSW distribution over time, providing a deeper understanding of temporal trends. Such approaches would enhance our ability to develop targeted and effective public health strategies for addressing the needs of vulnerable populations.

References

[1] R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/.

[2] Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686.

[3] Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Mächler, M., Bolker, B. M. (2017). glmmTMB Balances Speed and Flexibility Among Packages for Zero-Inflated Generalized Linear Mixed Modeling. *The R Journal*, *9*(2), 378–400. https://doi.org/10.32614/RJ-2017-066.

[4] Hartig, F. (2020). DHARMa: Residual Diagnostics for Hierarchical (Multi-Level/Mixed) Regression Models. Available at: https://CRAN.R-project.org/package=DHARMa.

[5] Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686.

[6] Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., Elberg, A., Crowley, J. (2024). GGally: Extension to 'ggplot2'. R package version 2.2.1. Available at: https://CRAN.R-project.org/package=GGally.

[7] Xie, Y. (2021). knitr: A General-Purpose Package for Dynamic Report Generation in R. Available at: https://CRAN.R-project.org/package=knitr.

[8] Greenwood, M. (n.d.). *Stat 505 Course Notes*. [Online].

[9] OpenAI. (n.d.). ChatGPT. [Online].

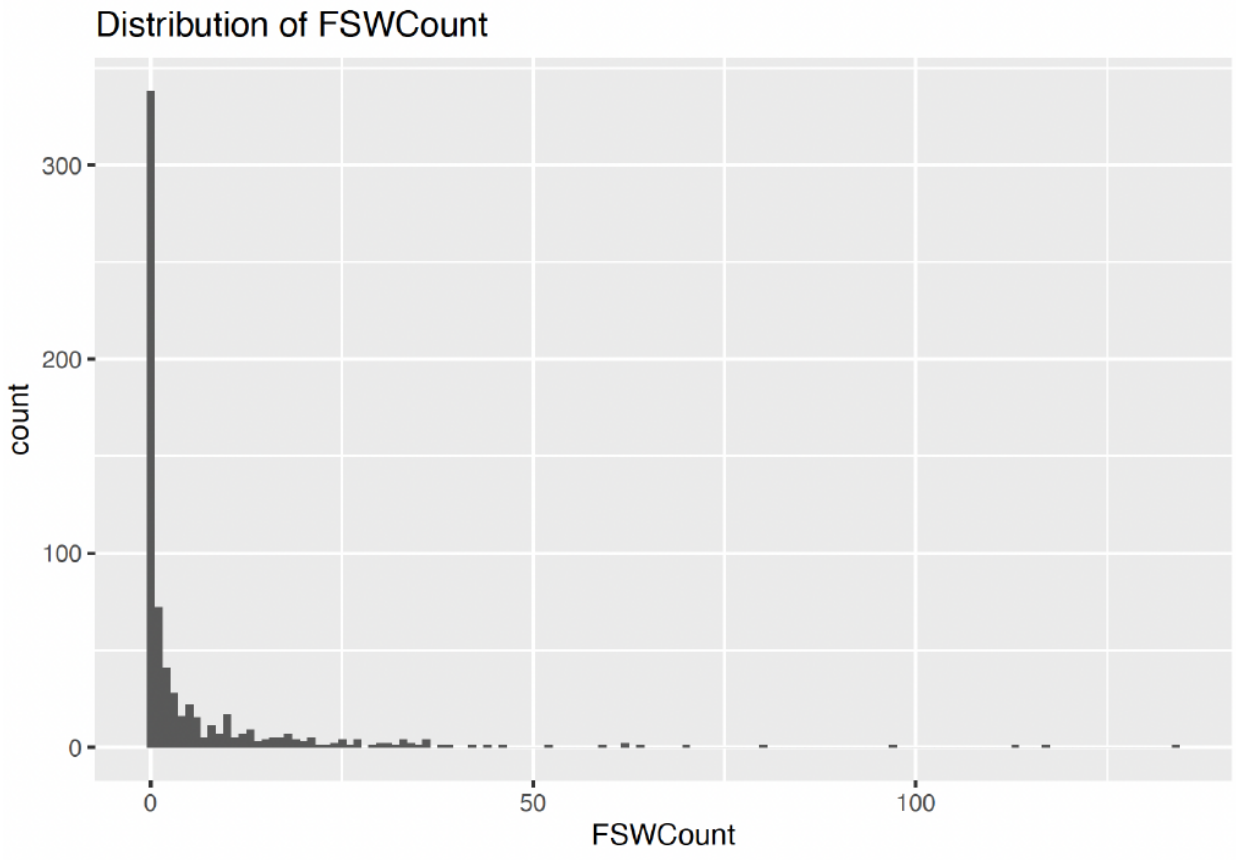

Figure 1: The distribution of female sex worker counts

Table 1: Table 3: AIC Scores for Mixed-Effect Models Using Environmental Variables

Model	**AIC**
Poisson	5978
Negative Binomial	2874
Hurdle Poisson	4900
Hurdle Negative Binomial	2999
Zero-inflated Poisson	2876
Zero-inflated Negative Binomial	4861

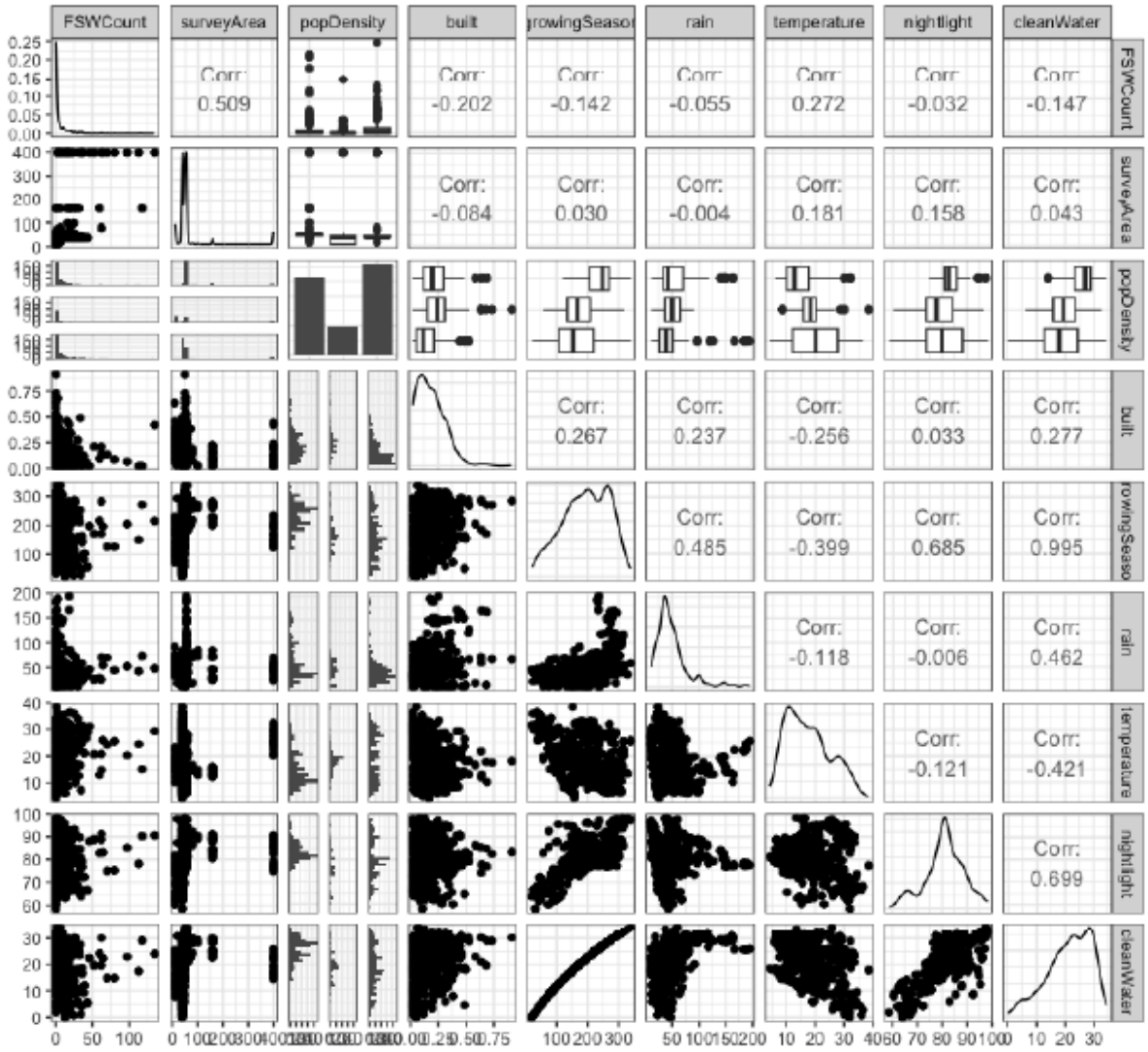

Figure 2: Correlation matrix of the variables showing the correlation between the variables.

Table 2: Table 4: AIC Scores for Mixed-Effect Models Using Sociological Variables

Model	**AIC**
Poisson	5551
Negative Binomial	2873
Hurdle Poisson	4602
Hurdle Negative Binomial	2989
Zero-inflated Poisson	2875
Zero-inflated Negative Binomial	4562

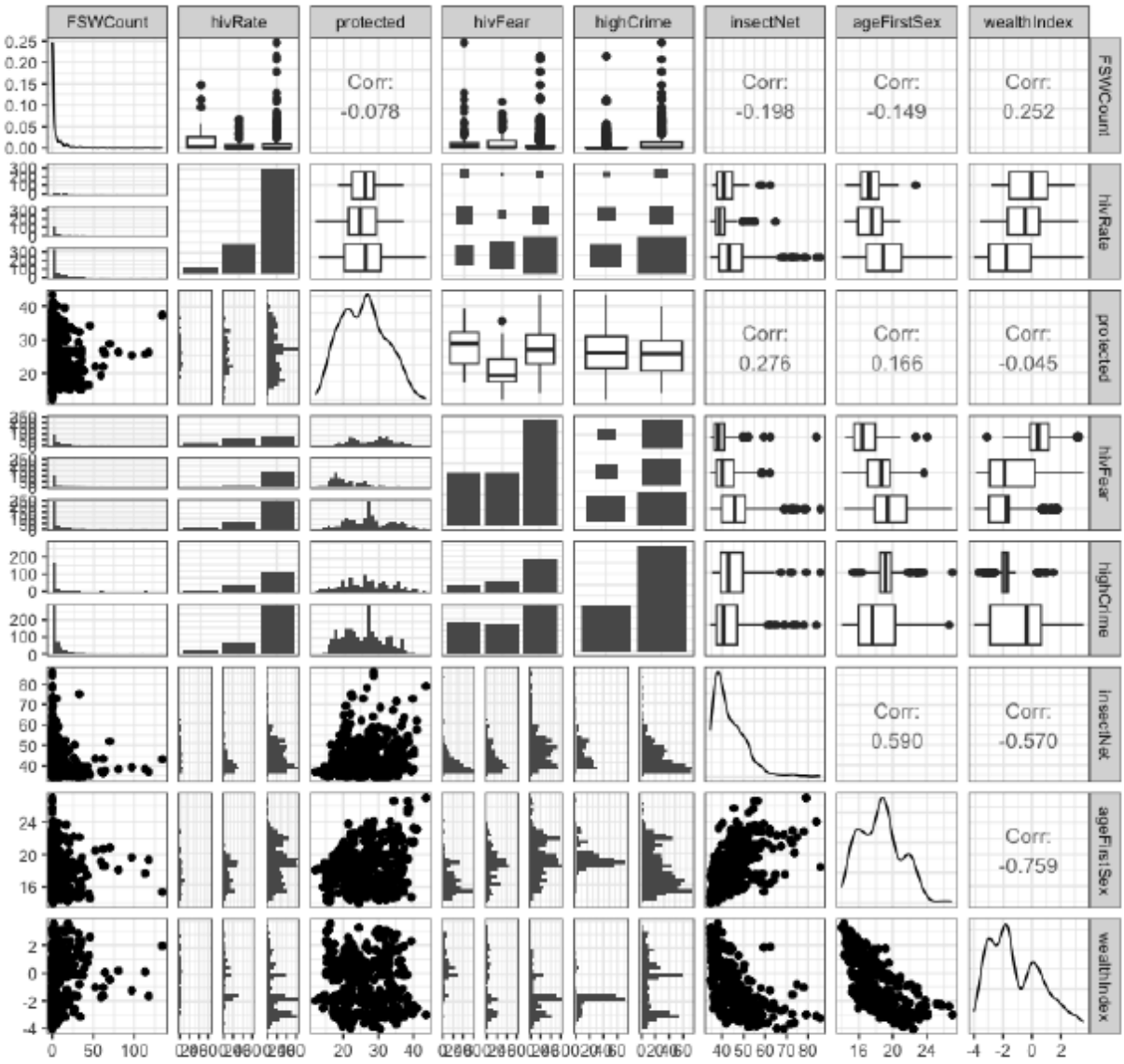

Figure 3: Correlation matrix of the variables showing the correlation between the variables

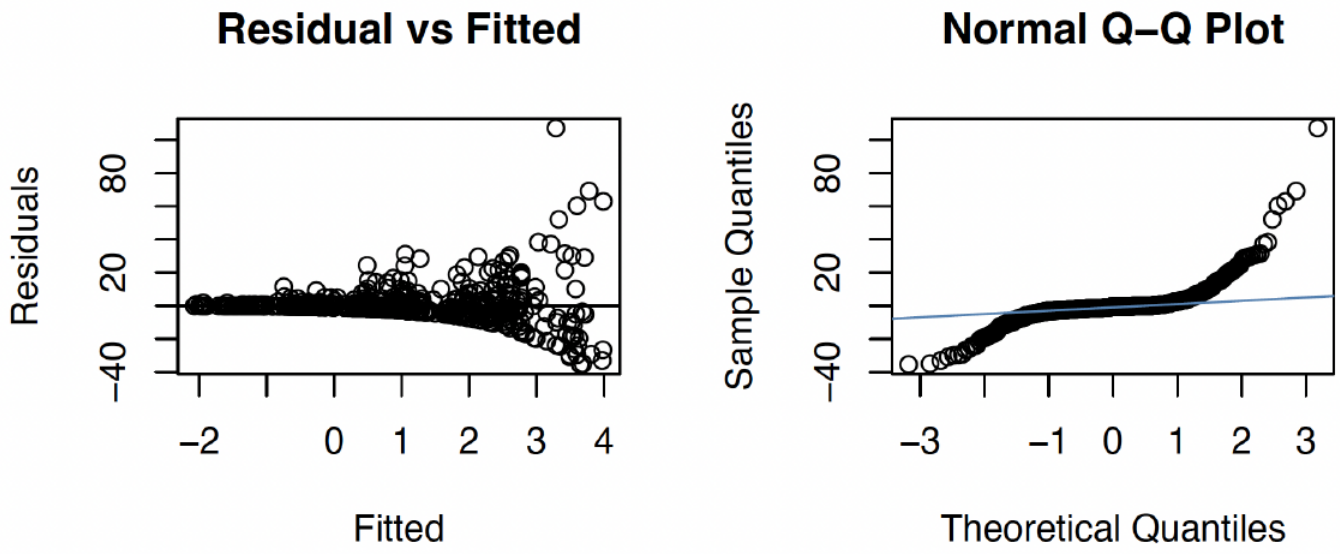

Figure 4: Diagnostic plot of the poisson mixed-effect model using environmental variables.

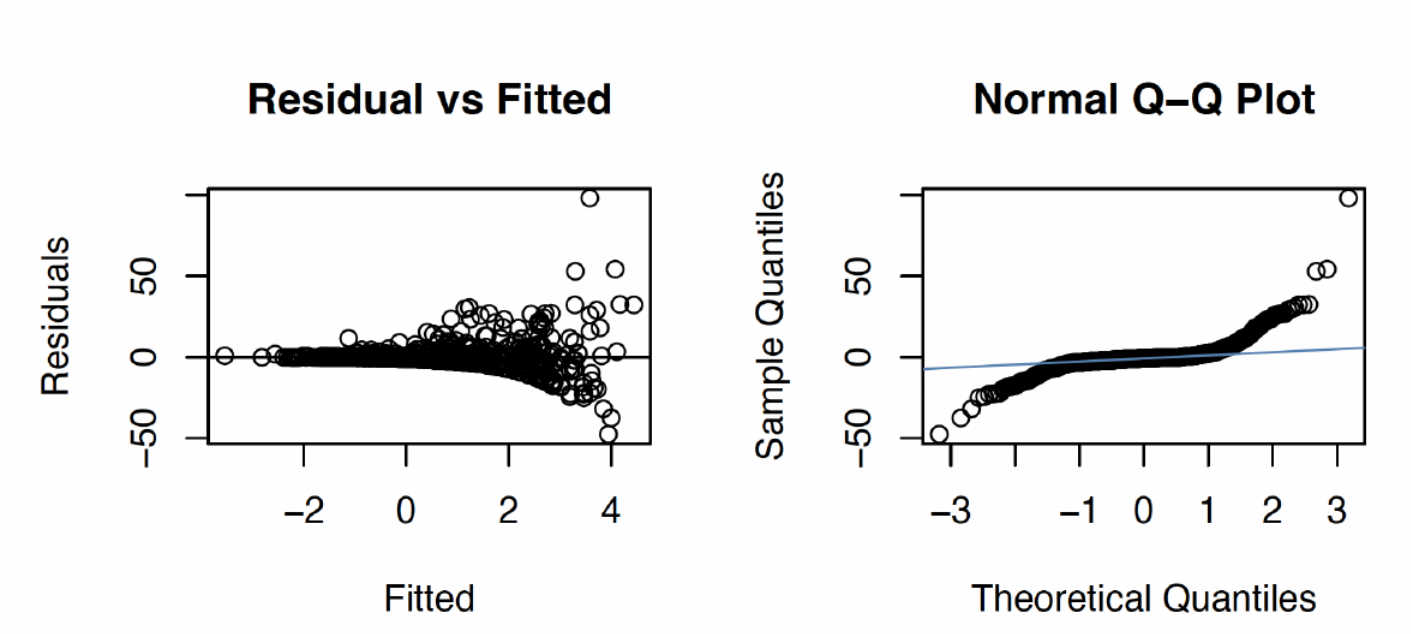

Figure 5: Diagnostic plot of the poisson mixed-effect model using sociological variables.

1 Appendix

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library(MASS)
library(GGally)
library(pscl)
library(pander)
library(lmerTest)
library(patchwork)

load.data <- read_csv("fsw_data.csv")
view(load.data)

training <- load.data %>%
  mutate(
    country = factor(country),
    region = factor(region),
    month = factor(month),
    dataYear = factor(dataYear),
    popDensity = factor(popDensity),
    hivRate = factor(hivRate),
    hivFear = factor(hivFear),
    highCrime = factor(highCrime)
  ) %>%
  drop_na()

testing <- load.data %>%
  mutate(
    country = factor(country),
    region = factor(region),
    month = factor(month),
    dataYear = factor(dataYear),
    popDensity = factor(popDensity),
    hivRate = factor(hivRate),
    hivFear = factor(hivFear),
    highCrime = factor(highCrime)
  ) %>%
  filter(is.na(FSWCount))
```

```
## Data Visualization and Exploration
library(GGally)

training %>%
  ggpairs(5:13,
          upper = list(continuous = GGally::wrap(ggally_cor, stars = FALSE))) +
  theme_bw()

training %>%
  ggpairs(c(5, 14:20),
          upper = list(continuous = GGally::wrap(ggally_cor, stars = FALSE))) +
  theme_bw()

echo = TRUE, warning = FALSE, message = FALSE, fig.cap = "This figure is a histogram plot of the Count

ggplot(training, aes(x = FSWCount)) +
  geom_histogram(binwidth = 10, fill = "lightblue", color = "black") +
  labs(x = "Female Sex Workers", y = "Frequency", title = "Histogram plot of the Count of FSW") +
  theme_minimal()
```

```
# MODEL SELECTION FOR ENVIRONMENTAL VARIABLES
library(pscl)
library(glmmTMB)

fit1 <- glmmTMB(FSWCount ~ popDensity + built + growingSeason + rain +
  temperature + nightlight + cleanWater + (1 | country) + (1 | region) +
  (1 | month) + (1 | dataYear), offset = log(surveyArea),
  data = training, family = poisson)

fit2 <- glmmTMB(FSWCount ~ popDensity + built + growingSeason + rain +
  temperature + nightlight + cleanWater + (1 | country) + (1 | region) +
  (1 | month) + (1 | dataYear), offset = log(surveyArea),
  data = training, family = nbinom2)

fit3 <- glmmTMB(FSWCount ~ popDensity + built + growingSeason + rain +
  temperature + nightlight + cleanWater + (1 | country) + (1 | region) +
  (1 | month) + (1 | dataYear), offset = log(surveyArea), zi = ~1,
  family = truncated_poisson, data = training)

fit4 <- glmmTMB(FSWCount ~ popDensity + built + growingSeason + rain +
  temperature + nightlight + cleanWater + (1 | country) + (1 | region) +
  (1 | month) + (1 | dataYear), offset = log(surveyArea),
  family = truncated_nbinom2, data = training, zi = ~1)
```

```
# RANDOM EFFECT SELECTION FOR ENVIRONMENTAL VARIABLES
fit2A <- glmmTMB(FSWCount ~ popDensity + built + growingSeason + rain +
  temperature + nightlight + cleanWater + (1 | country) + (1 | region) +
  (1 | month) + (1 | dataYear), offset = log(surveyArea),
  data = training, family = nbinom2, REML = TRUE)

fit2B <- glmmTMB(FSWCount ~ popDensity + built + growingSeason + rain +
  temperature + nightlight + cleanWater + (1 | country) + (1 | region) +
  (1 | month), offset = log(surveyArea),
  data = training, family = nbinom2, REML = TRUE)

fit2C <- glmmTMB(FSWCount ~ popDensity + built + growingSeason + rain +
  temperature + nightlight + cleanWater + (1 | country) + (1 | region) +
  (1 | dataYear), offset = log(surveyArea),
  data = training, family = nbinom2, REML = TRUE)
```

```
# FIXED VARIABLE SELECTION FOR ENVIRONMENTAL VARIABLES
fit2i <- glmmTMB(FSWCount ~ popDensity + built + growingSeason + rain +
  temperature + nightlight + cleanWater + (1 | country) + (1 | region),
  offset = log(surveyArea),
  data = training, family = nbinom2, REML = FALSE)

fit2t <- glmmTMB(FSWCount ~ popDensity + built +
  temperature + (1 | country) + (1 | region),
  offset = log(surveyArea),
  data = training, family = nbinom2, REML = FALSE)

fit2u <- glmmTMB(FSWCount ~ 1 + (1 | country) + (1 | region),
  offset = log(surveyArea),
  data = training, family = nbinom2, REML = FALSE)
```

```
# MODEL SELECTION FOR SOCIOLOGICAL VARIABLES
fits1 <- glmmTMB(FSWCount ~ hivRate + protected + hivFear + highCrime +
  insectNet + ageFirstSex + wealthIndex + (1 | country) + (1 | region) +
  (1 | month) + (1 | dataYear), offset = log(surveyArea),
  data = training, family = poisson)

fits2 <- glmmTMB(FSWCount ~ hivRate + protected + hivFear + highCrime +
  insectNet + ageFirstSex + wealthIndex + (1 | country) + (1 | region) +
  (1 | month) + (1 | dataYear), offset = log(surveyArea),
  data = training, family = nbinom2)
```