MATH11188 Statistical Research Skills: Assignment 3

E. Antypas, C.Giannikos, Y. Gu, S. Liu

The purpose of this report is to obtain an estimation of the scale of modern slavery in the UK in 2013. To achieve this, an exploratory analysis of the *Modern Slavery Dataset* is performed, given that the data have been collected through different resources and the victim coverage is only partial. The figure of 2428 victims, as recorded in the four given lists, contains only cases which were encountered and recorded by at least one of the organisations. As a result, it does not include the potential victims which were not encountered by any of the four. Therefore, the multiple systems estimation technique can be used to estimate the unknown number of cases that have not come to attention. The subsequent statistical analysis determines a suitable generalised linear model, in terms of the two-way interactions present within the model, and examines the absolute goodness-of-fit of the final selected model. The implementation of the required methods in *R* can be found in the *Appendix*.

1 Statistical Methodology

1.1 Model Selection

To model the data, a generalised linear model (GLM) is considered, from the Poisson family, with log-link function. The response variable corresponds to the number of individuals observed by each combination of sources; and the (discrete) explanatory variables are the corresponding lists. Because the lists are not assumed to be independent of each other, the model also considers the two-way interactions between the explanatory variables. Consequently, the number of cases not recorded by any of the four lists is given by

$$y_{0000}|\mu_{0000} \sim Poisson(\mu_{0000}),$$

where $\log(\mu_{0000}) = \beta_0$. Therefore, the estimate of the true total of potential victims, i.e. the quantity of interest, is given by

$$\hat{N} = exp(\hat{\beta}_0) + n.$$

In order to determine which of the interactions are statistically significant in describing the data, the Akaike Information Criterion (AIC) was used as a means of model selection(following (1)). This was achieved through the application of the (backward) **stepAIC** routine. The routine initially considered the full GLM, including the main effects and all the two-way interactions between them. It then proceeded by iteratively deducting interaction effects, at each stage removing the term which least least improved the AIC. The routine was terminated at the point where further removal would result in a worse fit according to the AIC. The model finally chosen for the four-list data contains all the main effects of the four lists and three of their six possible interactions, namely LA $\times$ NG, LA $\times$ PF and NG $\times$ GO. It must be noted that, complementary to the backward stepAIC, the forward stepAIC, as well as the stepAIC based on the BIC (Bayesian Information Criterion) were also applied and both yielded the exact same model.

1.2 Testing Goodness of Fit

The deviance of a GLM is an appropriate measure for goodness of fit, since it demonstrates the difference between the log-likelihood of said model and of a model that perfectly fits the data. Here, the deviance of the fitted model is 11.3 on 7 degrees of freedom. Furthermore, since the deviance D of the selected model follows a χ^2_7 distribution, the corresponding p-value for goodness-of-fit is given by the probability to the right of the deviance value for the chi-squared distribution on 7 degrees of freedom, i.e.

$$P(D < x) \approx 0.12 > 0.01,$$

and thus we do not have sufficient evidence to reject the null hypothesis H_0 : The model fits the data well at 0.01 level of significance. Additionally, in the Deviance Residual Plot for the model (Figure 1), the deviance residuals are evenly distributed around the x-axis, and do not demonstrate a systematic pattern, thus the goodness of fit hypothesis is not violated.

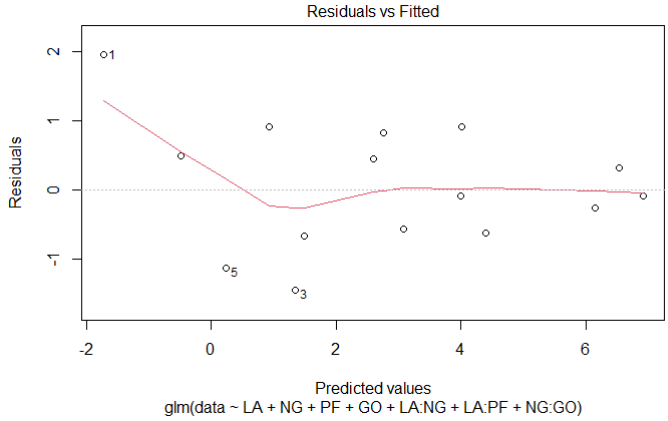

Figure 1: Deviance Residuals

1.3 Model Estimates

The model finally chosen for the four-list data provides an estimated total population size of 11015 with a standard error of 800. The respective estimate of the cases not reported in any of the lists is 8587, with the same standard error as the previous estimate. A 95% confidence interval is given by [9447, 12583]. Additionally, as demonstrated in the result output below, there is positive correlation between LA and each of NG and PF. This implies that being known to the local authority increases the chance of being known to NGOs or the police. However, there is a negative correlation between NG and GO, which implies that there exists a percentage of cases known to NGOs but unknown to Government agencies.

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.73059  -0.60592  -0.08187   0.62588   1.34699

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  9.05803    0.09315  97.242  < 2e-16 ***
LA          -5.05793    0.15418 -32.805  < 2e-16 ***
NG          -2.90609    0.09509 -30.563  < 2e-16 ***
PF          -2.14163    0.08810 -24.308  < 2e-16 ***
GO          -2.51454    0.09122 -27.564  < 2e-16 ***
LA:NG        1.49843    0.27687   5.412 6.23e-08 ***
LA:PF        0.89569    0.26283   3.408 0.000655 ***
NG:GO       -0.56185    0.22398  -2.508 0.012125 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 6047.262  on 14  degrees of freedom
Residual deviance:   11.341  on  7  degrees of freedom
AIC: 96.83

Number of Fisher Scoring iterations: 5
```

Figure 2: Selected Model Output

2 Conclusion

After implementing the multiple systems estimation technique in the given data regarding potential modern day slaves in the UK in 2013 and checking the goodness-of-fit for the proposed model, the total number of potential victims was estimated of being around 11015.

References

[1] B. W. Silverman, "Multiple-systems analysis for the quantification of modern slavery: classical and bayesian approaches," *Journal of the Royal Statistical Society Series A: Statistics in Society*, vol. 183, no. 3, pp. 691–736, 2020.

Appendix

```
################################### Appendix ###########################################
## This file contains the code used for Assignment 3 in Statistical Research
## Skills (MATH11188).
## Group Work:: E. Antypas (s2449453), S. Liu (s2337553),
## C. Giannikos (S2436019), Y. Gu(s2304572)



################################### Code ###############################################

## Creating dataset
data <- c(1, 1, 1, 15, 0, 3, 19, 54, 4, 19, 62, 464, 76, 703, 1006)
## Local Authority
LA <- c(1,1,1,1,1,1,1,1,0,0,0,0,0,0,0)
## Non-Government Organisation
NG <- c(1,1,1,1,0,0,0,0,1,1,1,1,0,0,0)
## Police Force
PF <- c(1,0,1,0,1,0,1,0,1,0,1,0,1,0,1)
## Government Organisation
GO <- c(1,1,0,0,1,1,0,0,1,1,0,0,1,1,0)

## Step-wise AIC procedure

## Model with all interactions
model.full <- glm(data~LA+NG+PF+GO
                  +LA:NG+LA:PF+LA:GO
                  +NG:PF+NG:GO+PF:GO, family=poisson(link="log"))

library(MASS)

## Model selection based on AIC (backward)
model.select <- stepAIC(model.full, direction="both")

## Summary
summary(model.select)

## model with no interactions
model.reduced <- glm(data~LA+NG+PF+GO, family=poisson(link="log"))

## Forward model selection based on AIC
stepAIC(model.reduced, scope=list(lower= formula(model.reduced),
upper=formula(model.full)),direction="forward")

## Model selection based on BIC
stepAIC(model.full, direction="both", criterion="BIC")

## P-value for goodness of fit
1-pchisq(summary(model.select)$deviance, summary(model.select)$df.residual)

## Residual plots
plot(model.select)

## Confidence Intervals
## Step 1: Extract the estimate and standard error of the intercept
coef <- summary(model.select)$coefficients[1,]
beta_0_hat<- coef["Estimate"]
```

```
beta_0_se <- coef["Std. Error"]

## Step 2: Calculate the estimate for N=exp(beta_0)+n
y_hat<- exp(beta_0_hat)
n <- sum(data)
N_hat <- y_hat+n

## Step 3: Calculate the standard error of N_hat
N_se <- y_hat* beta_0_se

## Step 4: Calculate the 95% confidence interval for y=exp(beta_0)
lb_ci <- N_hat - N_se*1.96
ub_ci <- N_hat + N_se*1.96
N_ci <- c(lb_ci,ub_ci)
```