

DAS Project2 Group18

Kejin Han, Yubin Lyu, Jiaxuan Miao, Chaoyi Zhang and Peihua Zhong

```
library(tidyverse)
library(dplyr)
library(readr)
library(ggplot2)
library(vcd)
library(MASS)
```

```
#load the data
shelter_01 <- read.csv("dataset18.csv")
```

```
#Checking for missing value
any_na <- apply(shelter_01, 2, function(x) any(is.na(x)))
any_na
```

animal_type	month	year	intake_type	outcome_type
FALSE	FALSE	FALSE	FALSE	FALSE
chip_status	time_at_shelter			
FALSE	FALSE			

```
total_na <- sum(is.na(shelter_01))
total_na
```

```
[1] 0
```

```
#Converting a string variable to a factor type and make a summary statistics
shelter_01$animal_type <- as.factor(shelter_01$animal_type)
shelter_01$intake_type <- as.factor(shelter_01$intake_type)
shelter_01$outcome_type <- as.factor(shelter_01$outcome_type)
```

```
shelter_01$chip_status <- as.factor(shelter_01$chip_status)
summary(shelter_01)
```

```

  animal_type    month      year      intake_type
BIRD      : 2   Min.    : 1.000   Min.    :2016   CONFISCATED    : 59
CAT       :238   1st Qu.: 4.000   1st Qu.:2017   OWNER SURRENDER:363
DOG       :880   Median : 7.000   Median :2017   STRAY          :713
LIVESTOCK: 1   Mean    : 6.574   Mean    :2017
WILDLIFE : 14   3rd Qu.: 9.000   3rd Qu.:2017
           Max.    :12.000   Max.    :2017

      outcome_type      chip_status  time_at_shelter
ADOPTION      :474   SCAN CHIP      :214   Min.      : 0.00
DIED          : 14   SCAN NO CHIP   :860   1st Qu.: 1.00
EUTHANIZED    :417   UNABLE TO SCAN: 61   Median : 4.00
FOSTER        : 30                                     Mean    : 6.12
RETURNED TO OWNER:200                                   3rd Qu.: 9.00
                                                         Max.    :78.00
```

```
#Converting shelter_01 to dataframe
shelter_02 <- as.data.frame(shelter_01)
summary(shelter_02)
```

```

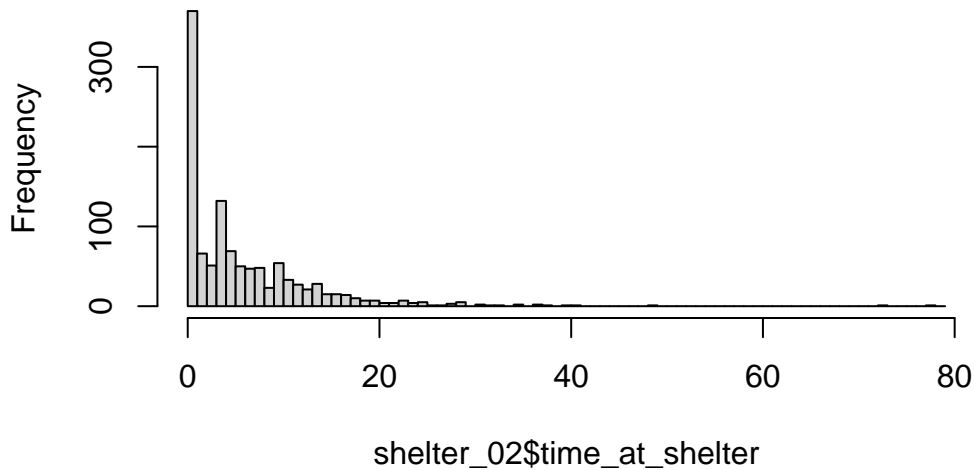
  animal_type    month      year      intake_type
BIRD      : 2   Min.    : 1.000   Min.    :2016   CONFISCATED    : 59
CAT       :238   1st Qu.: 4.000   1st Qu.:2017   OWNER SURRENDER:363
DOG       :880   Median : 7.000   Median :2017   STRAY          :713
LIVESTOCK: 1   Mean    : 6.574   Mean    :2017
WILDLIFE : 14   3rd Qu.: 9.000   3rd Qu.:2017
           Max.    :12.000   Max.    :2017

      outcome_type      chip_status  time_at_shelter
ADOPTION      :474   SCAN CHIP      :214   Min.      : 0.00
DIED          : 14   SCAN NO CHIP   :860   1st Qu.: 1.00
EUTHANIZED    :417   UNABLE TO SCAN: 61   Median : 4.00
FOSTER        : 30                                     Mean    : 6.12
RETURNED TO OWNER:200                                   3rd Qu.: 9.00
                                                         Max.    :78.00
```

```
# Histogram analysis of the dependent variable
hist(shelter_02$time_at_shelter,
```

```
breaks = seq(min(shelter_02$time_at_shelter),
             max(shelter_02$time_at_shelter) + 1, by = 1))
```

Histogram of shelter_02\$time_at_shelter



The significance of finding out whether the dependent variable is continuous or count is in choosing the appropriate statistical method and model for the analysis. If the histogram shows a continuous and smooth distribution, it usually indicates that the dependent variable is continuous. If the histogram shows a discrete and spaced distribution, it usually indicates that the dependent variable is of the count type.

From the results, it can be known that histogram is showing interval shape and overall is not smooth. Therefore the dependent variable is count type. Count variables are usually analysed using Poisson regression and negative binomial distribution regression. Therefore, we have attempted to use Poisson distribution and negative binomial distribution regression for the subsequent Generalized Linear Model respectively.

```
glm_model_poi <- glm(time_at_shelter ~ year + month + animal_type + intake_type
                    + outcome_type + chip_status, data = shelter_02,
                    family = poisson())
summary(glm_model_poi)
```

Call:

```
glm(formula = time_at_shelter ~ year + month + animal_type +  
      intake_type + outcome_type + chip_status, family = poisson(),  
      data = shelter_02)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	209.563215	219.298918	0.956	0.33927
year	-0.108790	0.043770	-2.486	0.01294 *
month	-0.016839	0.005842	-2.882	0.00395 **
animal_typeCAT	13.253664	200.734972	0.066	0.94736
animal_typeDOG	13.354757	200.734971	0.067	0.94696
animal_typeLIVESTOCK	-0.191216	348.317912	-0.001	0.99956
animal_typeWILDLIFE	12.834001	200.735017	0.064	0.94902
intake_typeOWNER SURRENDER	-1.367180	0.049511	-27.614	< 2e-16 ***
intake_typeSTRAY	-0.856870	0.044964	-19.057	< 2e-16 ***
outcome_typeDIED	-0.469573	0.113310	-4.144	3.41e-05 ***
outcome_typeEUTHANIZED	-0.542380	0.027585	-19.662	< 2e-16 ***
outcome_typeFOSTER	-0.576073	0.088272	-6.526	6.75e-11 ***
outcome_typeRETURNED TO OWNER	-1.621092	0.050170	-32.312	< 2e-16 ***
chip_statusSCAN NO CHIP	-0.258643	0.031581	-8.190	2.62e-16 ***
chip_statusUNABLE TO SCAN	-0.645688	0.074825	-8.629	< 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: 8495.8 on 1134 degrees of freedom
Residual deviance: 6544.7 on 1120 degrees of freedom
AIC: 9670.3

Number of Fisher Scoring iterations: 10

```
model_poi <- step(glm_model_poi)
```

Start: AIC=9670.31

```
time_at_shelter ~ year + month + animal_type + intake_type +  
      outcome_type + chip_status
```

	Df	Deviance	AIC
<none>		6544.7	9670.3

```
- year          1    6550.9  9674.5
- month         1    6553.0  9676.6
- animal_type   4    6587.0  9704.6
- chip_status    2    6651.2  9772.8
- intake_type    2    7270.5 10392.1
- outcome_type   4    8056.1 11173.7
```

The stepwise analysis of the model was carried out while performing the Poisson distribution. The initial model had an AIC value of 9670.31. In further steps, the independent variables such as year, month, animal_type, chip_status, intake_type, and outcome_type were gradually deleted but these deletion operations all lead to an increase in the AIC value, indicating that deleting these variables makes the model worse. Therefore, it can be seen that the best model is when no independent variables are added or removed, which corresponds to an AIC value of 9670.3 and a deviation of 6544.7.

```
glm_model_nb <- glm.nb(time_at_shelter ~ year + month + animal_type + intake_type
                        + outcome_type + chip_status, data = shelter_02)
summary(glm_model_nb)
```

Call:

```
glm.nb(formula = time_at_shelter ~ year + month + animal_type +
        intake_type + outcome_type + chip_status, data = shelter_02,
        init.theta = 0.9633756977, link = log)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	3.078e+02	1.333e+05	0.002	0.9982
year	-1.638e-01	1.217e-01	-1.345	0.1785
month	-2.029e-02	1.613e-02	-1.258	0.2084
animal_typeCAT	2.619e+01	1.333e+05	0.000	0.9998
animal_typeDOG	2.631e+01	1.333e+05	0.000	0.9998
animal_typeLIVESTOCK	-3.126e-01	2.315e+05	0.000	1.0000
animal_typeWILDLIFE	2.574e+01	1.333e+05	0.000	0.9998
intake_typeOWNER SURRENDER	-1.703e+00	1.600e-01	-10.640	< 2e-16 ***
intake_typeSTRAY	-1.295e+00	1.506e-01	-8.602	< 2e-16 ***
outcome_typeDIED	-4.871e-01	3.005e-01	-1.621	0.1050
outcome_typeEUTHANIZED	-6.033e-01	7.598e-02	-7.940	2.02e-15 ***
outcome_typeFOSTER	-4.783e-01	2.175e-01	-2.199	0.0279 *
outcome_typeRETURNED TO OWNER	-1.843e+00	1.108e-01	-16.638	< 2e-16 ***
chip_statusSCAN NO CHIP	-1.717e-01	9.032e-02	-1.901	0.0573 .
chip_statusUNABLE TO SCAN	-7.708e-01	1.816e-01	-4.244	2.20e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.9634) family taken to be 1)

Null deviance: 1640.1 on 1134 degrees of freedom
Residual deviance: 1312.5 on 1120 degrees of freedom
AIC: 6252.2

Number of Fisher Scoring iterations: 1

Theta: 0.9634
Std. Err.: 0.0542

2 x log-likelihood: -6220.2480

```
model_nb <- step(glm_model_nb)
```

Start: AIC=6250.25

time_at_shelter ~ year + month + animal_type + intake_type +
outcome_type + chip_status

	Df	Deviance	AIC
- month	1	1314.0	6249.7
- year	1	1314.3	6250.1
<none>		1312.5	6250.2
- animal_type	4	1325.0	6254.8
- chip_status	2	1330.3	6264.1
- intake_type	2	1439.1	6372.9
- outcome_type	4	1573.6	6503.4

Step: AIC=6249.74

time_at_shelter ~ year + animal_type + intake_type + outcome_type +
chip_status

	Df	Deviance	AIC
- year	1	1313.0	6248.2
<none>		1312.6	6249.7
- animal_type	4	1325.5	6254.6
- chip_status	2	1329.9	6263.0
- intake_type	2	1439.1	6372.2

```
- outcome_type 4 1574.3 6503.4
```

Step: AIC=6248.17

```
time_at_shelter ~ animal_type + intake_type + outcome_type +  
chip_status
```

	Df	Deviance	AIC
<none>		1312.6	6248.2
- animal_type	4	1325.4	6253.0
- chip_status	2	1330.1	6261.7
- intake_type	2	1439.7	6371.3
- outcome_type	4	1577.9	6505.5

We then tried the negative binomial distribution and performed a stepwise analysis of the model. The initial model had an AIC value of 6250.25. In the iteration, the year and month variables were gradually removed and each step resulted in a decrease in the AIC value. The final model contains independent variables such as `animal_type`, `intake_type`, `outcome_type` and `chip_status`.

By comparing the results of `model_poi` and `model_nb`, we can see the difference in the performance of the two models in the stepwise regression analysis. `model_nb` model obtained a lower AIC value by gradually deleting the year and month variables when selecting the variables, indicating that this model fits the data better, while `model_poi` model did not find that it could be further optimised during the stepwise regression process. `model_poi` model did not find any variables that could further optimise the model, so it retained all the independent variables, but with a relatively high AIC value.

By comparing the deviation values of the two models, `model_nb` has a final deviation value of 1312.6 and `model_poi` has a final deviation value of 6544.7. We can find that `model_nb` has a smaller deviation value and fits the data better than `model_poi`. Therefore, based on the comparison of deviation values, `model_nb` model is more suitable for interpreting and predicting the data.