Influential factors of days spent at the shelter

Hanfan Chen 2703334C, Zhaohao Li 2611602L, Zhenhao Qiao 2612824Q, Chao Wang 2518608W, Racha

1   Introduction

2   Exploratory Data Analysis

```
'data.frame':   1450 obs. of  7 variables:
 $ animal_type    : chr  "CAT" "DOG" "DOG" "DOG" ...
 $ month          : int  9 6 12 9 11 12 6 1 2 4 ...
 $ year           : int  2017 2017 2016 2017 2016 2016 2017 2017 2017 2017 ...
 $ intake_type    : chr  "STRAY" "STRAY" "STRAY" "STRAY" ...
 $ outcome_type   : chr  "ADOPTION" "EUTHANIZED" "ADOPTION" "ADOPTION" ...
 $ chip_status    : chr  "UNABLE TO SCAN" "SCAN NO CHIP" "SCAN NO CHIP" "SCAN NO CHIP" ...
 $ time_at_shelter: int  9 4 21 4 7 4 4 5 0 15 ...


Column 1 :
[1] "BIRD"     "CAT"       "DOG"       "WILDLIFE"


Column 2 :
 [1] "1"   "2"   "3"   "4"   "5"   "6"   "7"   "8"   "9"   "10" "11" "12"


Column 3 :
[1] "2016" "2017"


Column 4 :
[1] "CONFISCATED"     "OWNER SURRENDER" "STRAY"


Column 5 :
[1] "ADOPTION"        "DIED"              "EUTHANIZED"
[4] "FOSTER"          "RETURNED TO OWNER"


Column 6 :
[1] "SCAN CHIP"       "SCAN NO CHIP"    "UNABLE TO SCAN"
```

All the explanatory variable are categorical variable and each explanatory variable have multiple levels.

```
[1] 317
```

Over 300 zeros in raw data may cause overdispersion in Poisson regression. The hurdle model is suggested to fit.
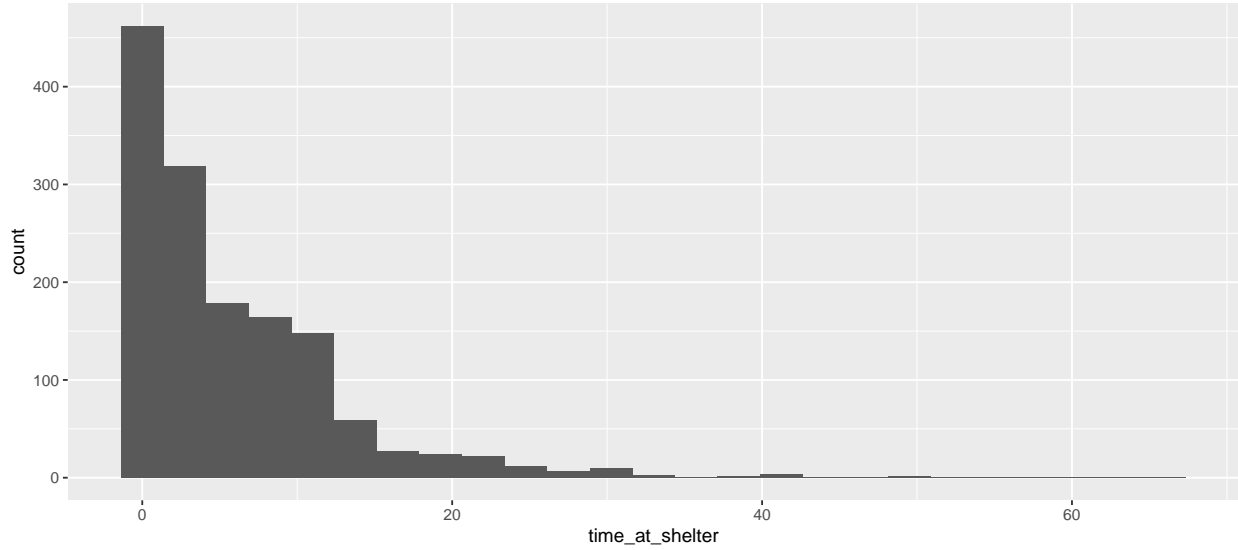
Figure 1: The histogram of day time at shelter

Figure 1 displays the histogram of the response variable, which is days time in the shelter. The histogram shows evidence of right-skewed and Poisson distribution of the response variable.

Figure 2: Pair plots of the variables

Note that year and month are strongly correlated, suggesting multicollinearity. As the data were collected over the period of a year, month and year represent the same variable, namely when the animal was admitted to the shelter. Therefore, year shall be omitted from the model.

The other explanatory variables are all categorical and their box plots are shown. The median time at shelter appears to be low for all the explanatory variables, which is due to the median time at shelter being 4.

Since the response variable is right-skewed, a median of the response variable is calculated. The figures below display the median of each category of the different explanatory variables.
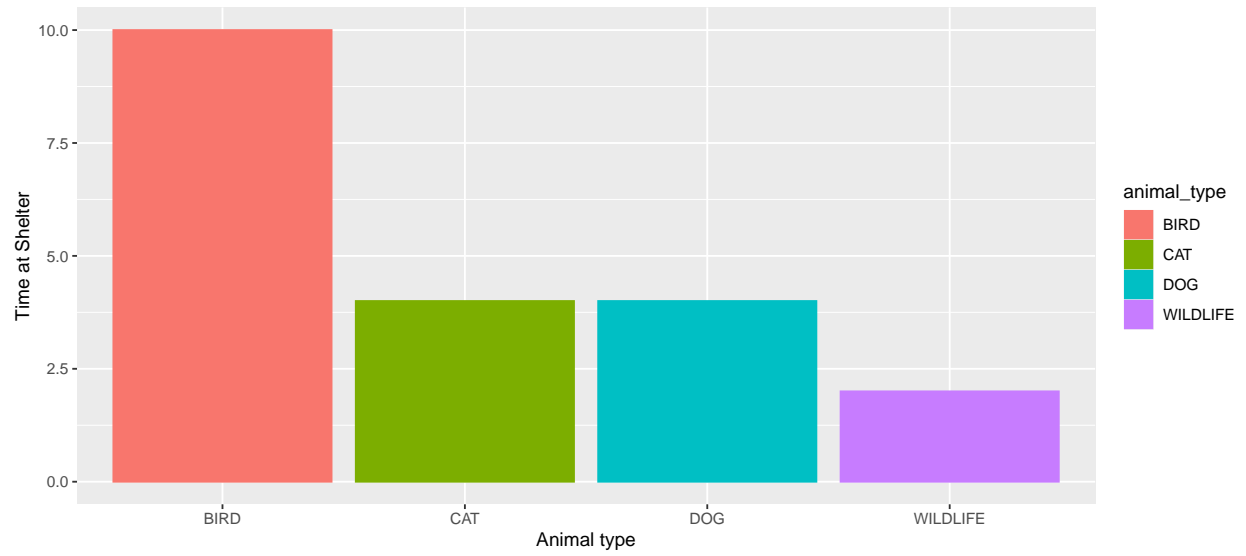
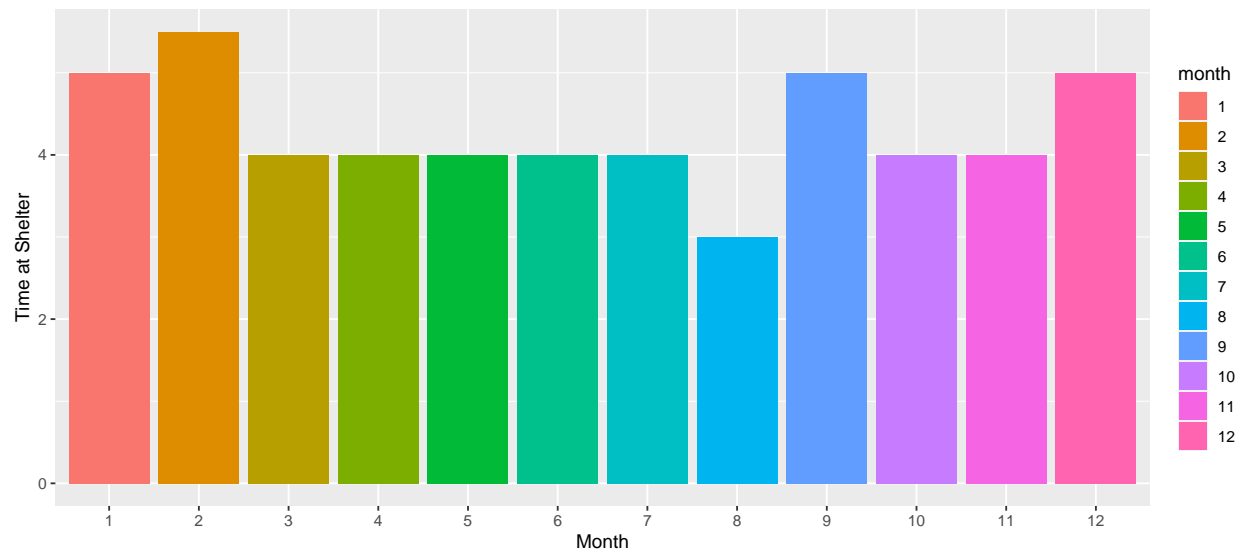Figure 3: Bar plot of animal type vs time at shelter

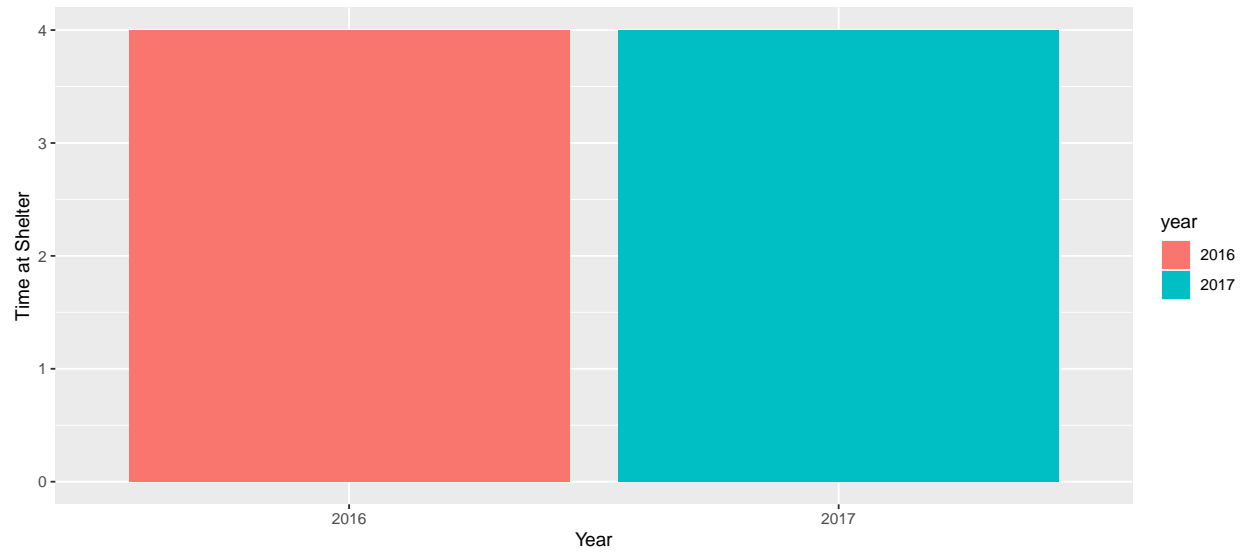Figure 4: Bar plot of month vs time at shelter

Figure 5: Bar plot of year vs time at shelter

`[1] FALSE FALSE FALSE`

No overlap between the months and years, according to the bar plot5, no obvious difference between two years and the relationship between the response variable and month variable is similar to that relationship between the response variable and the year variable. Therefore, the year variable is removed.

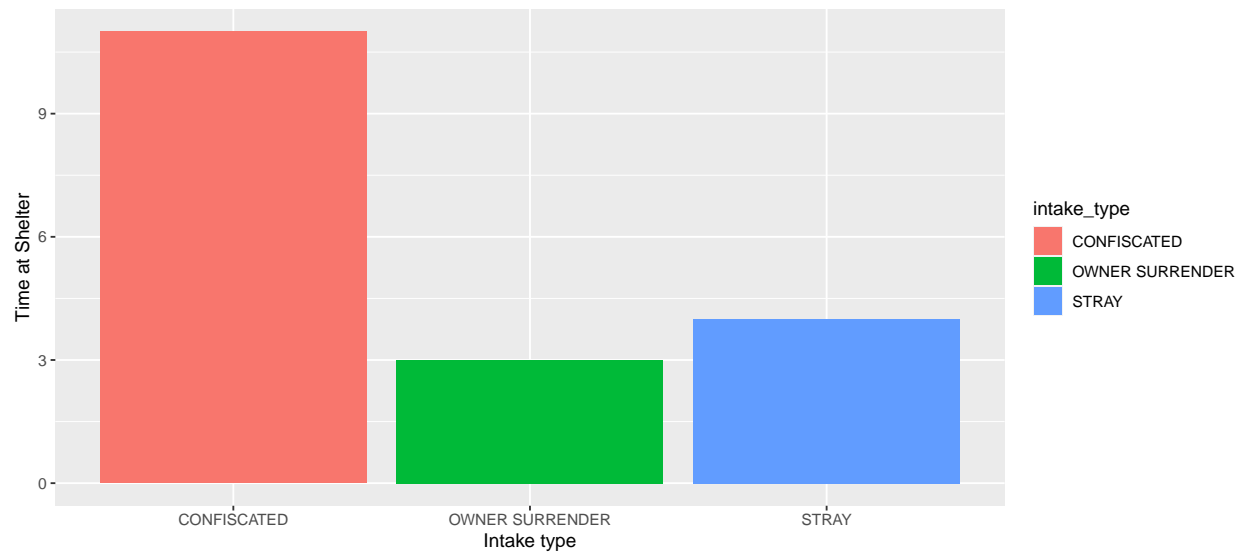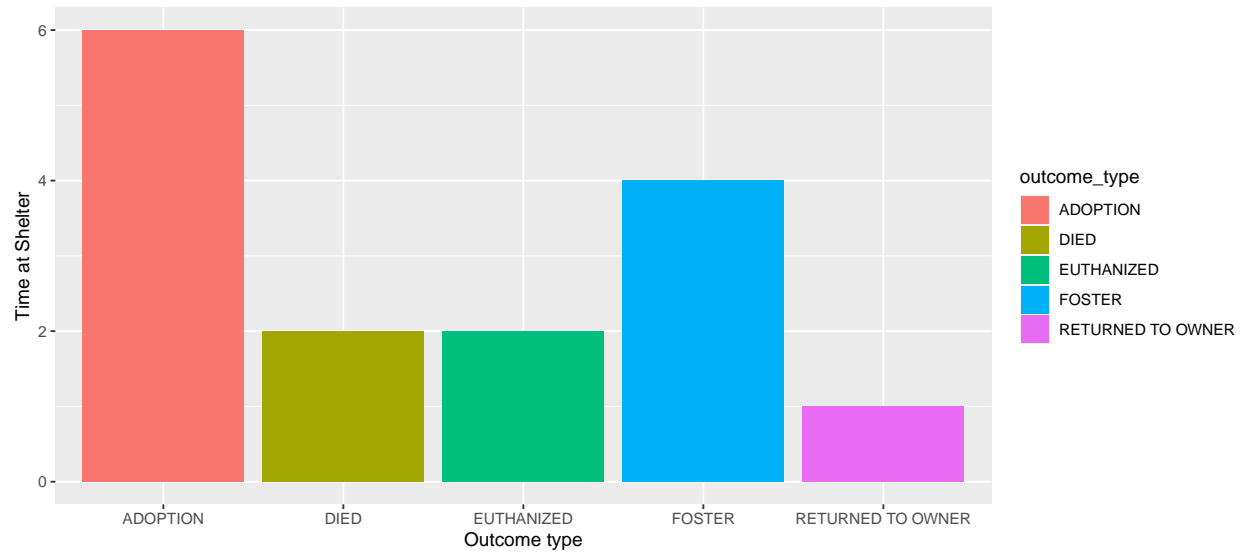Figure 6: Bar plot of intake type vs time at shelter

Figure 7: Bar plot of outcome type vs time at shelter
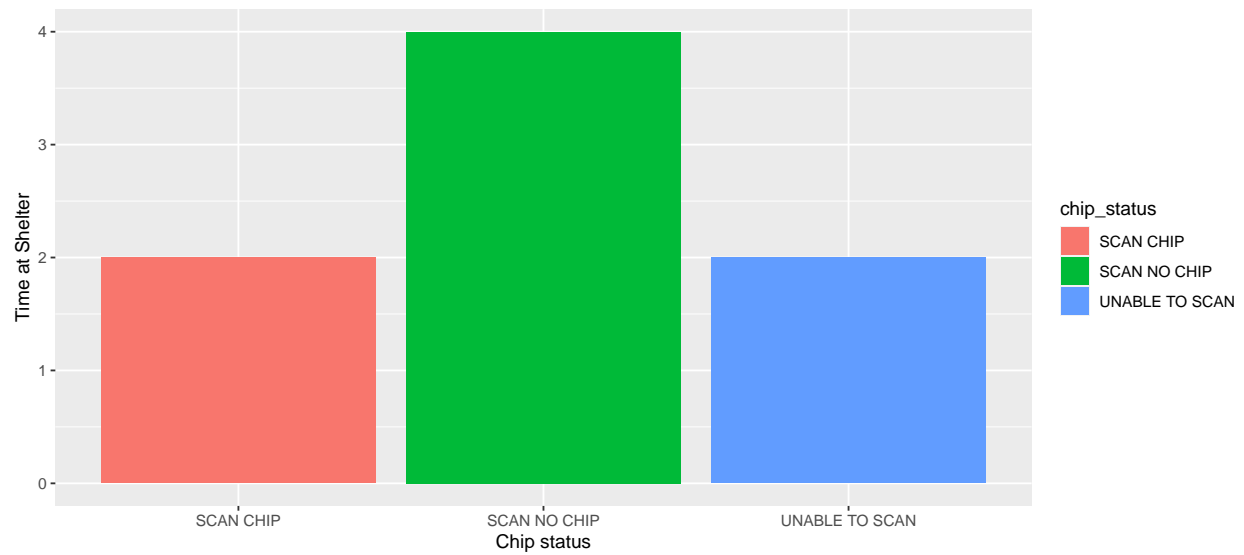

Figure 8: Bar plot of chip status vs time at shelter

3 Formal Data Analysis

3.1 Fitting a Poisson model

```
Call:
glm(formula = time_at_shelter ~ ., family = "poisson", data = data10)

Deviance Residuals:
    Min      1Q   Median      3Q      Max
-6.9146  -1.9976  -0.8903   0.6306  12.7550
```

```
Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                    2.997158   0.197263  15.194  < 2e-16 ***
animal_typeCAT                 0.441668   0.195885   2.255 0.024150 *
animal_typeDOG                 0.485824   0.194425   2.499 0.012462 *
animal_typeWILDLIFE            0.225305   0.231453   0.973 0.330336
month2                         0.075718   0.055370   1.367 0.171470
month3                        -0.132108   0.057115  -2.313 0.020721 *
month4                        -0.193819   0.056691  -3.419 0.000629 ***
month5                        -0.005919   0.052007  -0.114 0.909386
month6                        -0.035721   0.050097  -0.713 0.475818
month7                        -0.057427   0.050613  -1.135 0.256526
month8                        -0.413755   0.058842  -7.032 2.04e-12 ***
month9                        -0.082308   0.056140  -1.466 0.142617
month10                        0.101852   0.051801   1.966 0.049273 *
month11                       -0.055580   0.054389  -1.022 0.306833
month12                        0.114138   0.051633   2.211 0.027065 *
intake_typeOWNER SURRENDER    -1.451530   0.043649 -33.254  < 2e-16 ***
intake_typeSTRAY              -1.031365   0.039395 -26.180  < 2e-16 ***
outcome_typeDIED              -0.649881   0.097578  -6.660 2.74e-11 ***
outcome_typeEUTHANIZED        -0.592552   0.025262 -23.456  < 2e-16 ***
outcome_typeFOSTER            -0.279520   0.076201  -3.668 0.000244 ***
outcome_typeRETURNED TO OWNER -1.531722   0.042358 -36.161  < 2e-16 ***
chip_statusSCAN NO CHIP       -0.171716   0.028935  -5.934 2.95e-09 ***
chip_statusUNABLE TO SCAN     -0.247414   0.068726  -3.600 0.000318 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 10551.2  on 1449  degrees of freedom
Residual deviance:  8079.3  on 1427  degrees of freedom
AIC: 12147

Number of Fisher Scoring iterations: 6


[1] 41
```

The Poisson model is severely under-fitting zero counts. There were 317 zero counts observed in the data set but the model only fitted 41. Hence a hurdle model will be fitted to provide a better fit.

3.2 Fitting a Hurdle model

```
Call:
hurdle(formula = time_at_shelter ~ ., data = data10, dist = "poisson",
    zero.dist = "binomial")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-4.3608 -1.0287 -0.5823  0.4795 14.9926


Count model coefficients (truncated poisson with log link):
```

```
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  2.9579923  0.1983275  14.915  < 2e-16 ***
animal_typeCAT               0.3743137  0.1965591   1.904 0.056867 .
animal_typeDOG               0.3213099  0.1951832   1.646 0.099723 .
animal_typeWILDLIFE          0.4412799  0.2325810   1.897 0.057786 .
month2                      -0.0007866  0.0555725  -0.014 0.988706
month3                      -0.1913094  0.0574189  -3.332 0.000863 ***
month4                      -0.2968745  0.0570389  -5.205 1.94e-07 ***
month5                      -0.0358694  0.0522504  -0.686 0.492405
month6                      -0.1290100  0.0505296  -2.553 0.010675 *
month7                      -0.0908291  0.0508464  -1.786 0.074043 .
month8                      -0.3531232  0.0594007  -5.945 2.77e-09 ***
month9                      -0.1700644  0.0563869  -3.016 0.002561 **
month10                      0.0425144  0.0518410   0.820 0.412164
month11                     -0.0777278  0.0545280  -1.425 0.154023
month12                      0.0460268  0.0517740   0.889 0.374006
intake_typeOWNER SURRENDER  -1.1067328  0.0453104 -24.426  < 2e-16 ***
intake_typeSTRAY            -0.7609702  0.0407405 -18.678  < 2e-16 ***
outcome_typeDIED            -0.6233442  0.0998502  -6.243 4.30e-10 ***
outcome_typeEUTHANIZED      -0.2197569  0.0254704  -8.628  < 2e-16 ***
outcome_typeFOSTER          -0.1110361  0.0769153  -1.444 0.148847
outcome_typeRETURNED TO OWNER -0.9857031  0.0450846 -21.863  < 2e-16 ***
chip_statusSCAN NO CHIP     -0.2019465  0.0290236  -6.958 3.45e-12 ***
chip_statusUNABLE TO SCAN   -0.2152199  0.0686741  -3.134 0.001725 **
Zero hurdle model coefficients (binomial with logit link):
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  1.905e+01  6.099e+02   0.031    0.975
animal_typeCAT              -1.328e+01  6.099e+02  -0.022    0.983
animal_typeDOG             -1.266e+01  6.099e+02  -0.021    0.983
animal_typeWILDLIFE        -1.454e+01  6.099e+02  -0.024    0.981
month2                      7.990e-01  4.898e-01   1.631    0.103
month3                      3.817e-01  4.040e-01   0.945    0.345
month4                      3.724e-01  4.020e-01   0.926    0.354
month5                     -9.406e-04  3.735e-01  -0.003    0.998
month6                      4.541e-01  3.702e-01   1.227    0.220
month7                      1.809e-01  3.643e-01   0.497    0.620
month8                     -2.548e-01  3.782e-01  -0.674    0.500
month9                      3.331e-01  3.984e-01   0.836    0.403
month10                     3.409e-01  3.981e-01   0.856    0.392
month11                     5.129e-02  4.062e-01   0.126    0.900
month12                     4.482e-01  4.345e-01   1.032    0.302
intake_typeOWNER SURRENDER -3.171e+00  5.161e-01  -6.143 8.07e-10 ***
intake_typeSTRAY           -2.406e+00  4.857e-01  -4.955 7.25e-07 ***
outcome_typeDIED           -8.929e-01  8.223e-01  -1.086    0.278
outcome_typeEUTHANIZED     -2.999e+00  2.661e-01 -11.273  < 2e-16 ***
outcome_typeFOSTER         -2.137e+00  5.383e-01  -3.969 7.21e-05 ***
outcome_typeRETURNED TO OWNER -4.203e+00  3.115e-01 -13.491  < 2e-16 ***
chip_statusSCAN NO CHIP    -1.024e-01  1.978e-01  -0.518    0.605
chip_statusUNABLE TO SCAN  -6.084e-01  3.793e-01  -1.604    0.109
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 30
Log-likelihood: -5193 on 46 Df
```
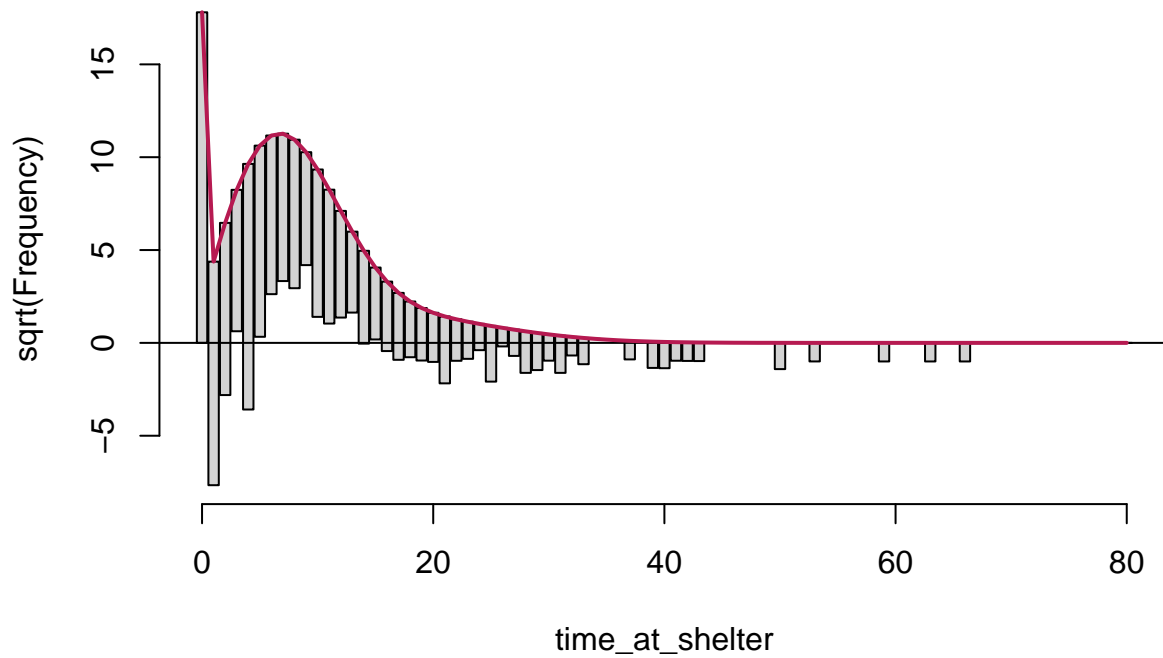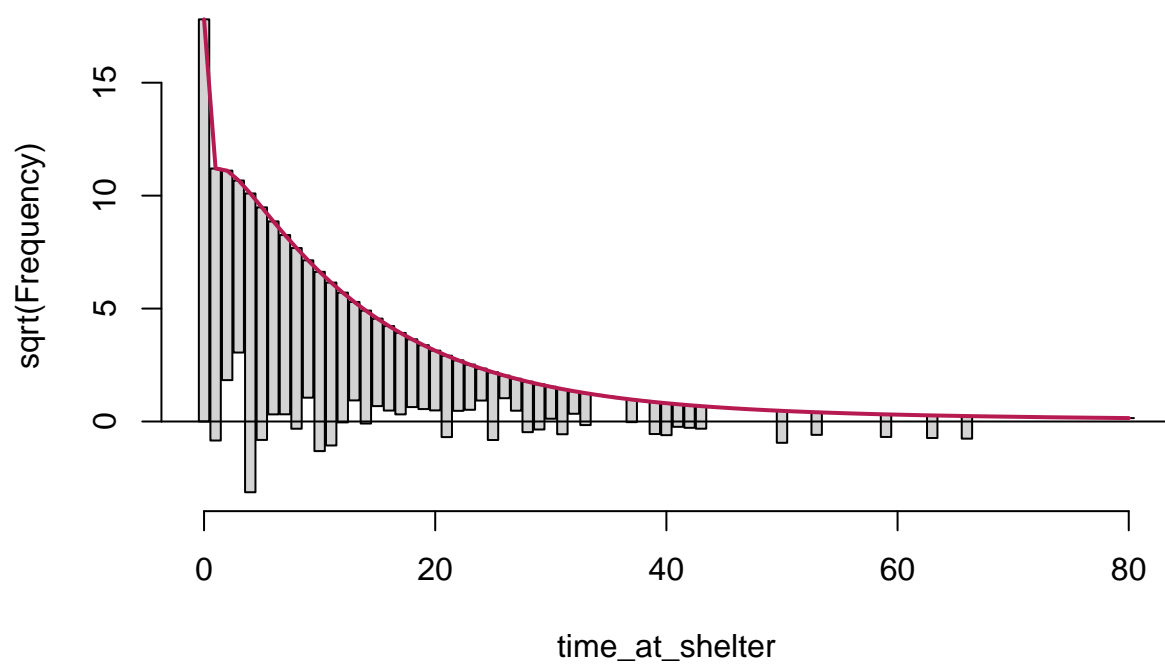
hurdle_model

The model is fitting the zero counts perfectly. However, counts 1,2 and 4 are being severely under-fitted, while 6-9 are being over-fitted. There is also under-fitting at the higher counts which suggests over-dispersion. Therefore, a negative binomial hurdle model shall be fitted to address this.

hurdle_model_nb

This shows a much better fit to the data. However, some values are still being under-fitted.

4 Conclusions