52542 Generalized Linear Models: Theory and Application - Final Paper

Omer Brandes - 302910476

March 4, 2018

A brief description of the problem:

We study counts of rat sightings in the city of Madrid. The brown rat lives with mankind and adversely affects public health by transmission of diseases, bites and allergies. Better understanding behavioural and spatial corre- lation aspects of this species can contribute to its effective management and control. We explore weakly to moderately correlated covariates based on distances to broken sewers, feeding grounds and markets as well as population density. The data were collected in Madrid city. It has a municipal surface area of about 605 km2 and 3.2 million inhabitants. In the context of large urban settlements, approximately 3% of the households have rats in their immediate environment, e.g.in compost heaps, gardens or unsecured rubbish bins. In the city of Madrid, direct sightings of rats and/or cockroaches or signs of their presence (e.g. droppings, burrows, gnaw marks, etc.) can be reported by citizens to the Technical Unit for Vector Control (TUVC). Only reports from people who declare to have sighted themselves any kind of these pest animals or their vital sign(s) in areas falling within the administrative borders of the municipality of Madrid are accepted. Records of the location and time of observation are entered in a dedicated database. The data used in this study contain the locations and dates of 6693 validated rat sightings reported to the TUVC from 1 January 2010 to 31 December 2013. [1]

Exploratory analysis of the data:

We have the following variables in the data set:
id, total.count, rat.count, ckr.count, xc, yc, market.dist, sewer.dist, catfeeding.dist

We will not use the id in the analysis since it obviously can't help us predict the response variable. In addition, since 'total.count' is just the sum of 'rat.count' and 'ckr.count', and therefore doesn't add any new information, we will not use it in the analysis as well. Some of the data:

Summary of the variables that will be used in the analysis:

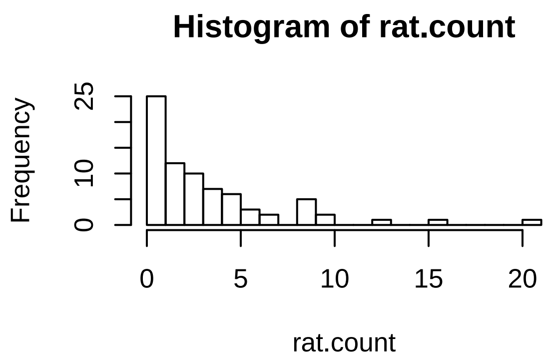
```
##    rat.count        ckr.count       market.dist      sewer.dist
##  Min.   : 0.000   Min.   : 0.000   Min.   :112.0   Min.   : 32.35
##  1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.:216.2   1st Qu.: 79.72
##  Median : 3.000   Median : 3.000   Median :370.8   Median :128.98
##  Mean   : 3.653   Mean   : 5.093   Mean   :386.2   Mean   :161.86
##  3rd Qu.: 5.000   3rd Qu.: 8.000   3rd Qu.:509.1   3rd Qu.:220.08
##  Max.   :21.000   Max.   :27.000   Max.   :979.9   Max.   :655.00
##  catfeeding.dist
##  Min.   : 29.40
##  1st Qu.: 91.47
##  Median :148.83
##  Mean   :184.54
##  3rd Qu.:210.40
##  Max.   :869.75
```

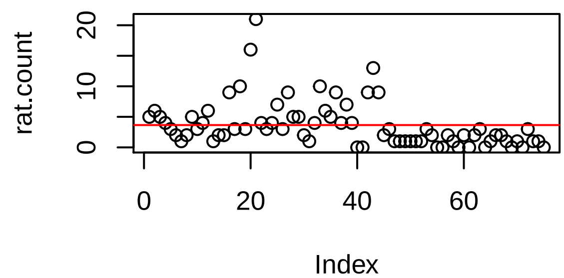
[1] Gräler, Benedikt, Carlos Ayyad, and Jorge Mateu. "Modelling count data based on weakly dependent spatial covariates using a copula approach: application to rat sightings." Environmental and Ecological Statistics 24, no. 3 (2017): 433-448.

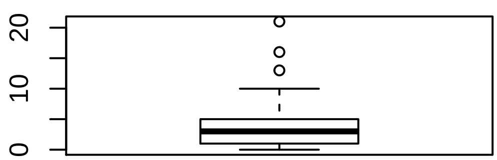
Exploring each variable:

The red line in the plots represents the mean value.

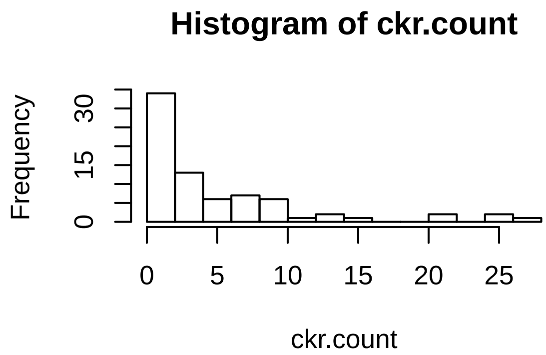
rat.count. We can see in the Histogram below that we have a large number of zeros, we will try to deal with this when we fit the models later on:

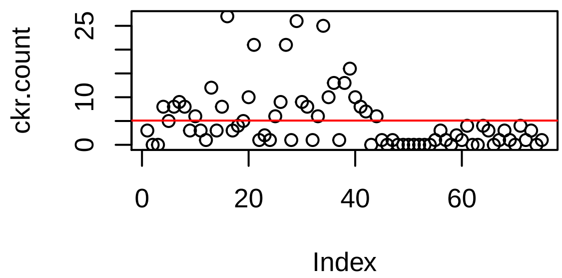


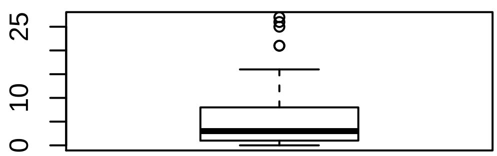




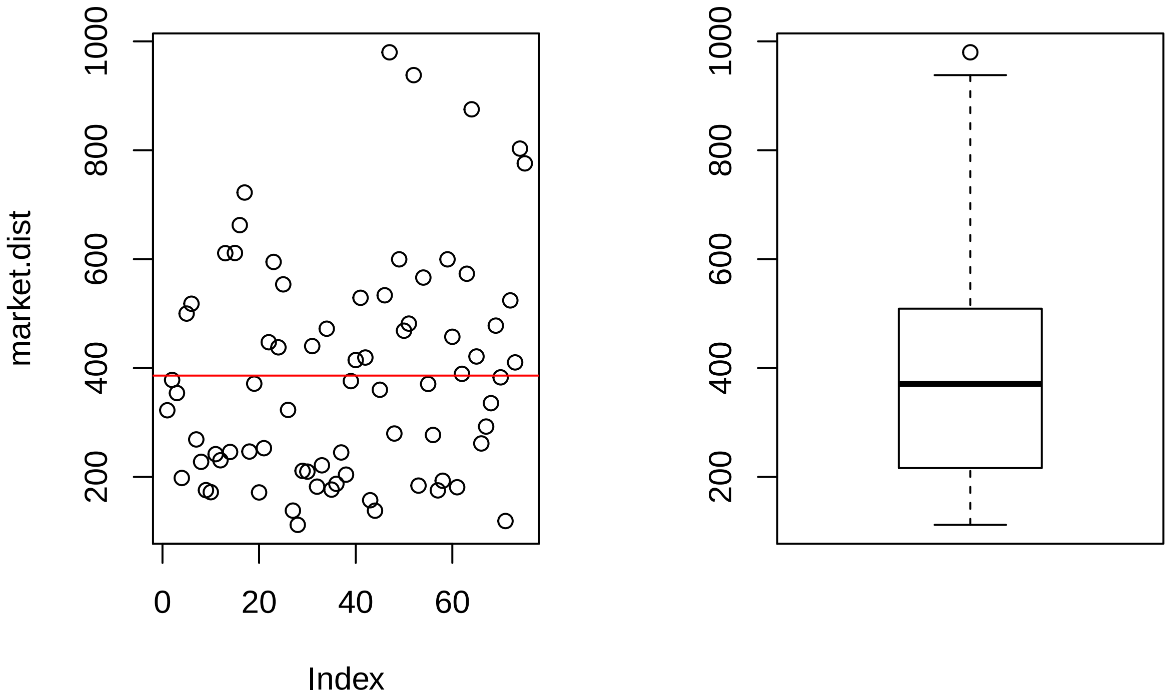
ckr.count:


Histogram of ckr.count
Frequency
0
15
30
0
5
10
15
20
25
ckr.count

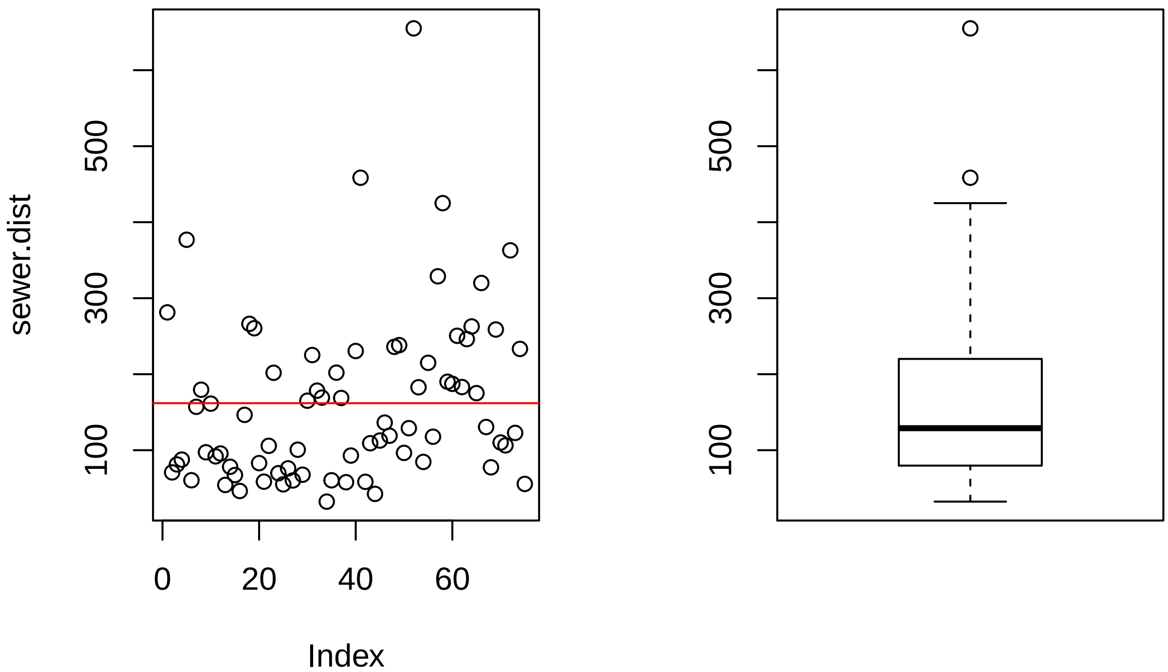

ckr.count
0
10
25
0
20
40
60
Index


0
10
25

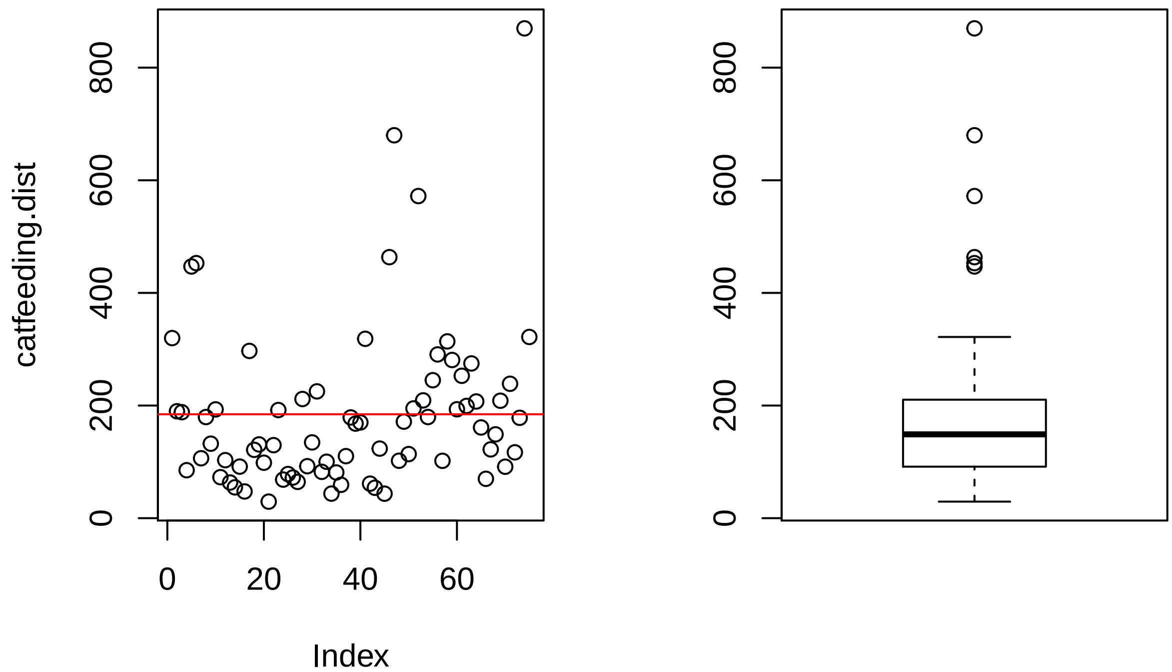
market.dist:


market.dist
Index
0
20
40
60
200
400
600
800
1000
200
400
600
800
1000

sewer.dist:



catfeeding.dist:



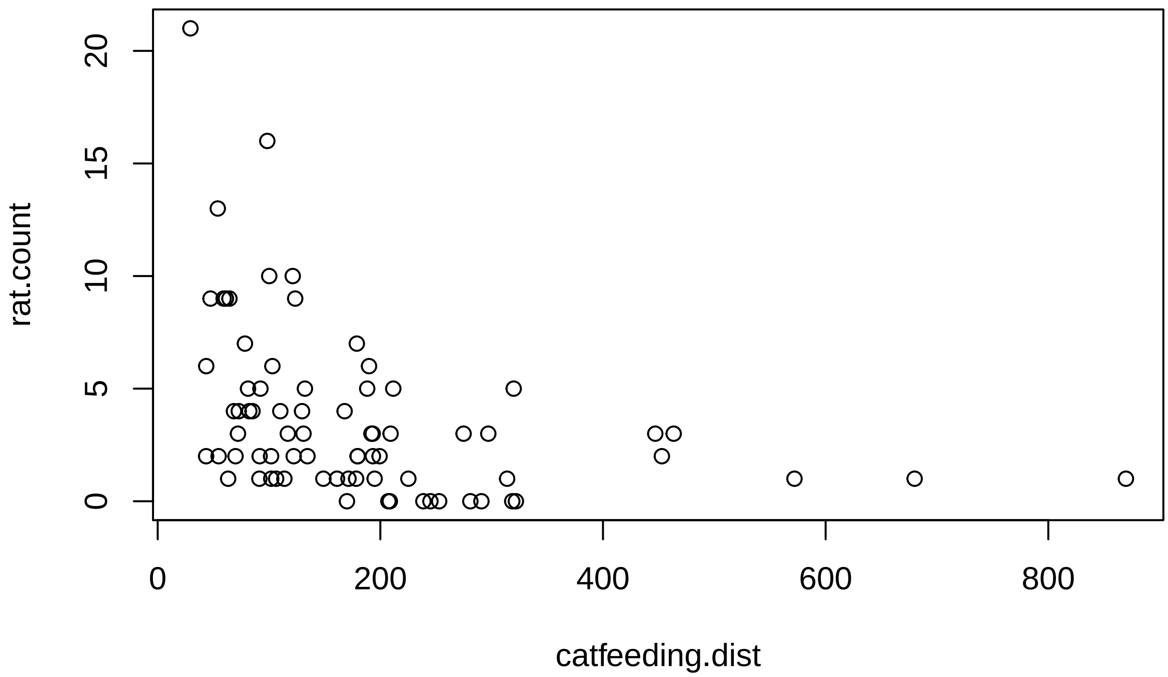
we have a few large observations in rat.count, we will try to see if these are outliers that should be removed from the dataset. We have three rat.count observation larger than two time the standard deviation of rat.count:

```
bigCount = which(rat.count > mean(rat.count)+2*sd(rat.count))

rats[bigCount,-c(2,5,6)]
```

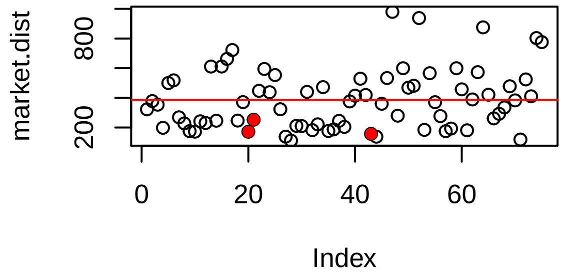
```
## # A tibble: 3 x 6
##      id rat.count ckr.count market.dist sewer.dist catfeeding.dist
##   <dbl>     <dbl>     <dbl>       <dbl>      <dbl>           <dbl>
## 1   495        16        10    171.5109   83.01087        98.46196
## 2   497        21        21    252.8182   58.59740        29.40260
## 3   924        13         0    157.2414  109.17241        53.96552
```

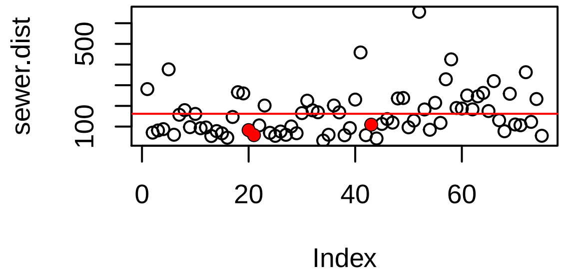
We expect the above observation to have below average distance from markets and sewers, but we first need to figure out wether cat feeding stations encourage rats or not:

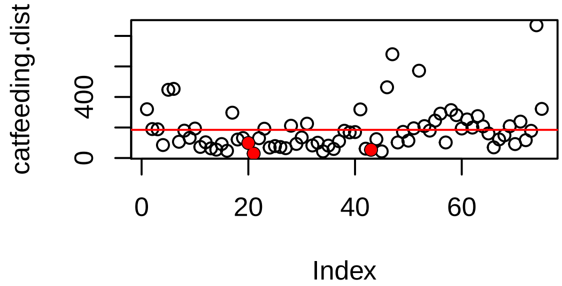


We can see that the negative trend in the above plot suggests that cat feeding stations encourage rats. Therefore, we expect that the possible outliers will also have below average cat feeding distance.

The possible outliers are marked with full red dots, the red horizontal line represents the mean value of the variable:

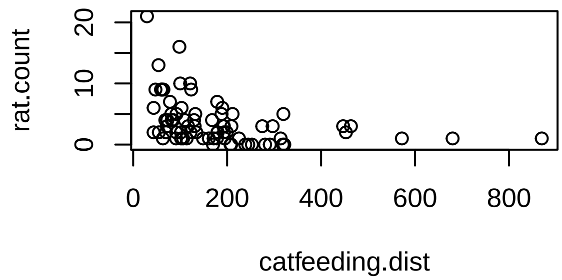


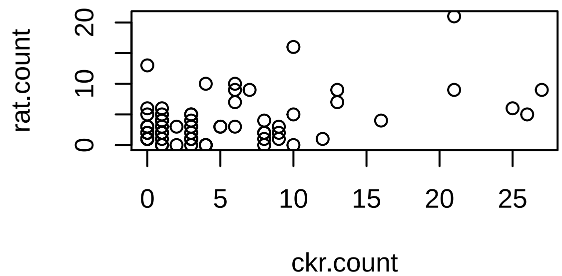


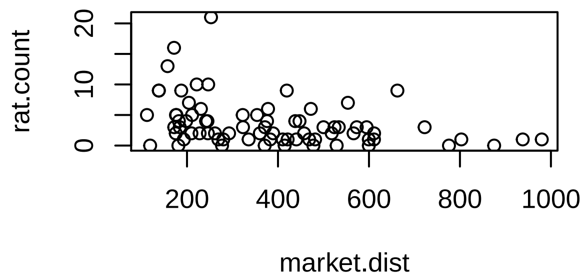


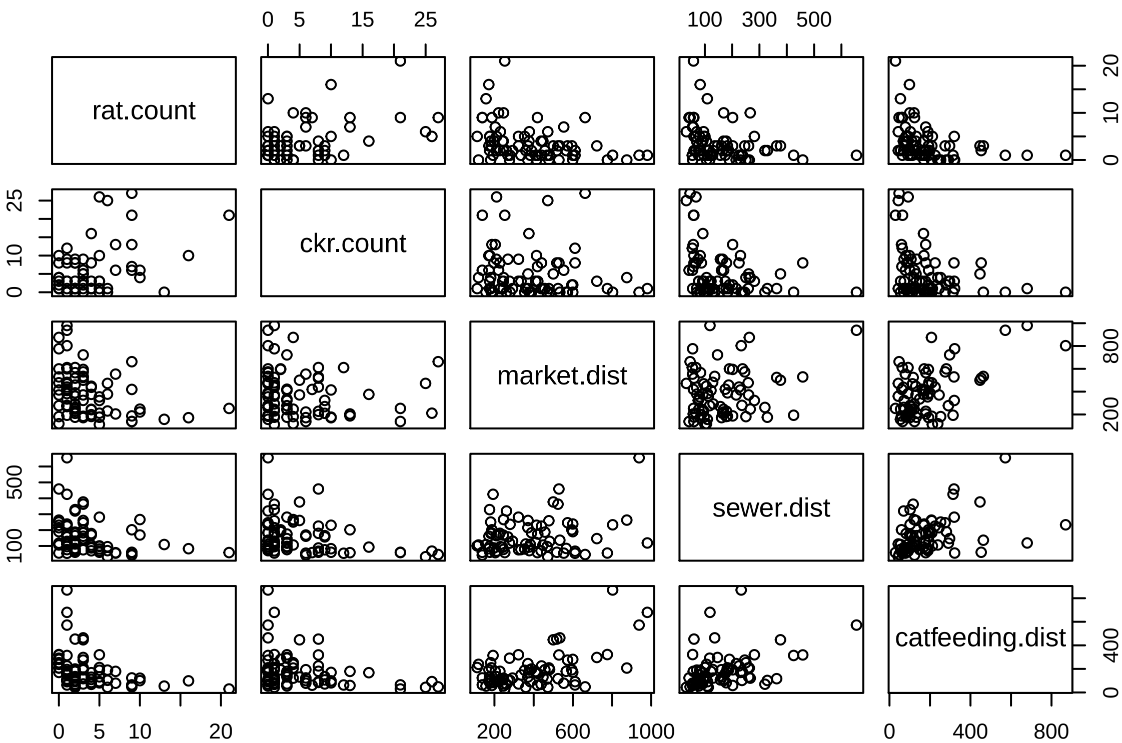
And we can see that the outliers are below average on all three variables, and therefore seems like we should keep them in the dataset.

plotting some graphs for further visualisation:


rat.count
0
10
20
0
200
400
600
800
catfeeding.dist


rat.count
0
10
20
0
5
10
15
20
25
ckr.count


rat.count
0
10
20
200
400
600
800
1000
market.dist



Covariance matrix - most are very low, with some excpeitons like rat.count/ckr.count and total.count, but that is expected:

```
##                         id total.count  rat.count  ckr.count          xc
## id               1.0000000  -0.4834610 -0.4440178 -0.3991121 -0.16820207
## total.count     -0.4834610   1.0000000  0.7591502  0.9209041  0.10457073
## rat.count       -0.4440178   0.7591502  1.0000000  0.4453846 -0.13711765
## ckr.count       -0.3991121   0.9209041  0.4453846  1.0000000  0.22594837
## xc              -0.1682021   0.1045707 -0.1371177  0.2259484  1.00000000
## yc              -0.3199761  -0.2155370 -0.1403726 -0.2124131  0.15990314
## market.dist      0.2309386  -0.2548710 -0.3544969 -0.1382928  0.11362922
## sewer.dist       0.2863353  -0.3848917 -0.3300170 -0.3317967 -0.18102155
## catfeeding.dist  0.2531637  -0.3915490 -0.3747576 -0.3141618  0.01892758
##                          yc market.dist  sewer.dist catfeeding.dist
## id              -0.31997610   0.2309386  0.28633533      0.25316372
## total.count     -0.21553697  -0.2548710 -0.38489175     -0.39154902
## rat.count       -0.14037265  -0.3544969 -0.33001702     -0.37475755
## ckr.count       -0.21241310  -0.1382928 -0.33179671     -0.31416176
## xc               0.15990314   0.1136292 -0.18102155      0.01892758
## yc               1.00000000   0.1233037  0.08115555      0.17211922
## market.dist      0.12330372   1.0000000  0.22301994      0.55536837
## sewer.dist       0.08115555   0.2230199  1.00000000      0.40540050
## catfeeding.dist  0.17211922   0.5553684  0.40540050      1.00000000
```

Model fitting:

We will start by fitting a Linear Model to the dataset. Since our response variable (rat.count) is a count variable, we don't expect the LM to fit well:

```
reg1 = lm(rat.count ~ ckr.count + market.dist + sewer.dist + catfeeding.dist)
reg1_summary = summary(reg1)
reg1_summary
```

```
##
## Call:
## lm(formula = rat.count ~ ckr.count + market.dist + sewer.dist +
##     catfeeding.dist)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7243 -2.0282 -0.4638  1.4878 12.6780
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.440410   1.075635   5.058 3.27e-06 ***
## ckr.count        0.205605   0.064522   3.187  0.00215 **
## market.dist     -0.004378   0.002276  -1.924  0.05841 .
## sewer.dist      -0.004480   0.003861  -1.160  0.24985
## catfeeding.dist -0.002267   0.003405  -0.666  0.50771
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.28 on 70 degrees of freedom
## Multiple R-squared:  0.3097, Adjusted R-squared:  0.2703
## F-statistic: 7.851 on 4 and 70 DF,  p-value: 2.751e-05
```

And indeed the model shows a very low R-squared value: 0.3097051.

Next, we will try to fit a poisson model:

```
glim1 = glm(rat.count ~ ckr.count + market.dist + sewer.dist + catfeeding.dist,
            family = "poisson", data = rats)
summary(glim1)
```

```
##
## Call:
## glm(formula = rat.count ~ ckr.count + market.dist + sewer.dist +
##     catfeeding.dist, family = "poisson", data = rats)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9165  -1.2901  -0.4266   0.9601   3.0098
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      2.2702279  0.2078716  10.921  < 2e-16 ***
## ckr.count        0.0306287  0.0085415   3.586 0.000336 ***
## market.dist     -0.0015975  0.0003960  -4.034 5.49e-05 ***
## sewer.dist      -0.0020007  0.0008764  -2.283 0.022448 *
## catfeeding.dist -0.0022722  0.0008408  -2.702 0.006882 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 255.50  on 74  degrees of freedom
## Residual deviance: 153.95  on 70  degrees of freedom
## AIC: 360.08
##
## Number of Fisher Scoring iterations: 5
```

We now attempt to add spatial variables to the model and see if there is a spatial trend, we fit a few models with different spatial variables such as x,y,x^2 ,y^2 , and compare their AIC and Devience:

```
cbind(glimSpatial1$aic, glimSpatial2$aic,glimSpatial3$aic,glimSpatial4$aic)
```

```
##          [,1]     [,2]     [,3]     [,4]
## [1,] 356.5614 362.0539 358.5461 358.5461
```

```
cbind(glimSpatial1$deviance, glimSpatial2$deviance,glimSpatial3$deviance,glimSpatial4$deviance)
```

```
##         [,1]     [,2]     [,3]     [,4]
## [1,] 148.438 153.9306 148.4228 148.4228
```

And finally, we compare the models with and without the sptial data:

```
cbind(glimSpatial1$aic, glim1$aic)
```

```
##          [,1]     [,2]
## [1,] 356.5614 360.0762
```

```
cbind(glimSpatial1$deviance, glim1$deviance)
```

```
##         [,1]     [,2]
## [1,] 148.438 153.9529
```

We will evaluate the poisson model with the spatial data using the Goodnes of Fit test: The GOF test indicates that the Poisson model doesn't fit the data ($p < 0.05$).

```
1 - pchisq(summary(glimSpatial1)$deviance, summary(glimSpatial1)$df.residual)
```

```
## [1] 9.532641e-08
```

This is expected because it seems like we have over-dispersion - the variance is much larger than the Expected Value (using the mean as an estimate to the Expected Value):

```
mean(rat.count)
```

```
## [1] 3.653333
```

```
var(rat.count)
```

```
## [1] 14.74306
```

In addition, we need to find a model that accounts for the zero inflation we noted earlier.

Negative-Binomial model:

```
glimNB = glm.nb(rat.count ~ ckr.count + market.dist + sewer.dist + catfeeding.dist,
                data = rats)
```

this model leads to an AIC of glimNB$aic. The devience is: 83.6183117 The Residual Sum of Squares: 670.1371289

```r
glimNB_spatial = glm.nb(rat.count ~ ckr.count + market.dist + sewer.dist + catfeeding.dist + xc,
                         data = rats)
```

Adding the spatial data of the x coordinates yields AIC glimNB_spatial$aic.Devience: glimNB_spatial$deviance. The Residual Sum of Squares is better: 629.2793005

So far NB model with spatial data gives the best aic and devience.

```r
rbind((cbind(glimNB$aic,glimNB$deviance)),
      (cbind(glimNB_spatial$aic,glimNB_spatial$deviance )))
```

```
##          [,1]     [,2]
## [1,] 338.0051 83.61831
## [2,] 337.2305 83.49150
```

Goodness of fit for NB models. The GOF test indicates that the Poisson model fits the data ($p > 0.05$).

```r
1 - pchisq(summary(glimNB_spatial)$deviance,
           summary(glimNB_spatial)$df.residual)
```

```
## [1] 0.1126906
```

Now we attempt to improve the model using a stepwise process:

```r
step(glimNB_spatial,direction="both")
```

```
## Start:  AIC=335.23
## rat.count ~ ckr.count + market.dist + sewer.dist + catfeeding.dist +
##     xc
##
##                   Df Deviance    AIC
## - catfeeding.dist  1   85.163 334.90
## <none>                 83.492 335.23
## - xc               1   86.320 336.06
## - sewer.dist       1   87.270 337.01
## - market.dist      1   89.450 339.19
## - ckr.count        1   91.522 341.26
##
## Step:  AIC=334.85
## rat.count ~ ckr.count + market.dist + sewer.dist + xc
##
##                   Df Deviance    AIC
## <none>                 82.497 334.85
## + catfeeding.dist  1   80.931 335.28
## - xc               1   85.645 336.00
## - sewer.dist       1   88.322 338.67
## - ckr.count        1   93.183 343.53
## - market.dist      1   93.230 343.58
```

```
##
## Call:  glm.nb(formula = rat.count ~ ckr.count + market.dist + sewer.dist +
##     xc, data = rats, init.theta = 3.579657749, link = log)
##
## Coefficients:
## (Intercept)    ckr.count  market.dist   sewer.dist           xc
##   2.044e+01    4.391e-02   -1.699e-03   -2.493e-03   -4.184e-05
##
## Degrees of Freedom: 74 Total (i.e. Null);  70 Residual
```

```
## Null Deviance:       130.6
## Residual Deviance: 82.5  AIC: 336.8
```

Seems like we can remove the catfeeding.dist variable.

```
glimNB_spatial_2 = glm.nb(rat.count ~ ckr.count + market.dist + sewer.dist + xc,
                          data = rats)
```

Goodness of fit improved: 0.1457715

AIC: 336.8473948 Devience: 82.4968401 Residual sum of squares: 667.3520599

In conclusion, we fitted the Negative-Binomial model to the data, and improved it using a stepwise process. The graph below shows the response variable (black empty dots) and the fitted valued (red full dots):

```
par(mfrow = c(1,1))
plot(rat.count)
points(glimNB_spatial_2$fitted.values,col='red', pch = 19)
```