GLM Final Paper - Rate Data

Ziv Parchi

Description

The final paper should analyze the data of rat sites in Madrid and conclude where (and why) there are rats in each pixel in Madrid. A pixel is 250m x 250m and there are 1,907 pixels dividing the city.
The variables described in our data were recorded for each pixel in Madrid during the period 1st of January 2010 to 31th of December 2013.
In our data we have the number of rats and cockroaches in each pixel (id). Furthermore, we have the data of the distance from the sampled pixel to the closest market, sewer and the distance from a cat feeding point, existence of cockroaches in the specific pixel, all parameters who can reasonably affect the existence of rats in a specific area. In addition, we have the x and y coordinates of each pixel, these are our spatial variables.
I got a sample of 95 different locations (pixels) of rat sites in Madrid and my goal is to analyze this data, while concluding the effect of each variable, and try to build a model which predicts the presence of rats in a given location alongside the relevant variables.
For this purpose, I would first extract basic important knowledge out of the data and try to get a better understanding of the relations between the different variables.

Variable	Description	N	Mean	SD	Median	Min	Max
rat.count	number of rats	95	3.694737e	4.538426	2.0000	0	25.000
ckr.count	number of cockroaches	95	3.947368e	4.675351	2	0	23.000
market.dist	the distance to the nearest market	95	4.833310e+02	735.606991	323.1304	8.070000e+01	7058.241
sewer.dist	the distance to the nearest sewer	95	1.851326e+02	219.999395	121.6667	3.180000e+01	1764.126
catfeeding.dist	the distance to the nearest cat-feeding station	95	2.331959e+02	273.278360	133.4595	3.288462e+01	1764.126

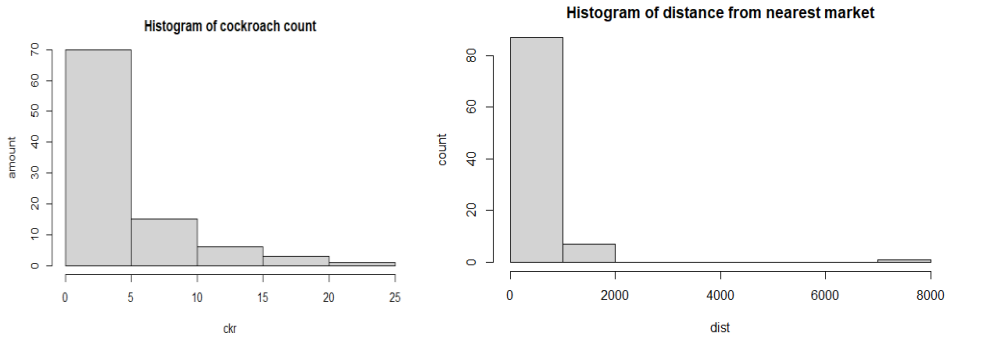
Preliminary exploratory analysis

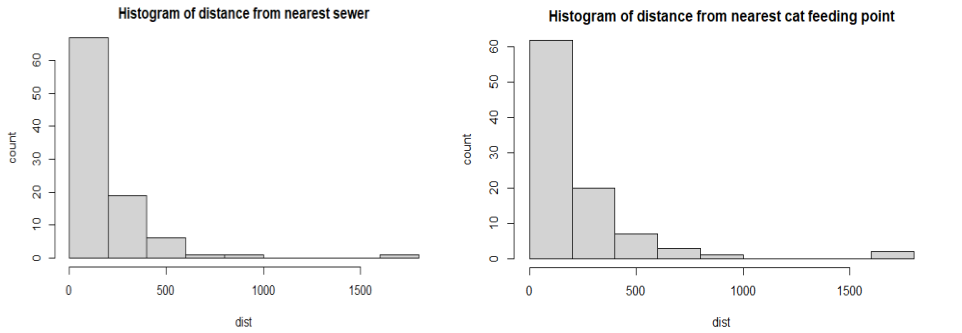
The explained variable Y is the ***rat.count*** variable, because given all the other explanatory variables we want to predict how many rats would be in a specified pixel. I would not use the *total.count* variable but i will use the *ckr.count* as an explanatory variable (due to the possible relation between the presence of cockroaches and rats at the same places).
In addition, it is easy to see that the id variable presented in our given data has no contribution for our goal so we should exclude it from our analysis. The reason is that the id variable does not make sense since the pixels are in 2-dimensional space.
It initially seems that there is no connection between the x,y coordinates and our explained variable, though, we might use these variables to find a connection between close pixels by finding spatial dependency.

Furthermore, it seems that the explained variable Y is referred to as the number of rats observed in the pixel over a specific time period (from our data description). So, it seems that it is reasonable to assume (after proper justifications) that we should be using a Poisson distributed model.

We would begin by fitting a linear model to the data in addition to a fit of a linear model with a log transformation, by calculating the explained variable as the log of the number of rats. Note that we would get inf for pixels where there are no rats, then we should add `1` inside the log. Now we get a new data frame('log_rat_data1'), in which the "rat.count" column/variable is now log(rat.count + 1).
Then, if necessary, we would fit a Poisson regression model since it seems that it would fit our data more reasonably than the linear model. Though, we might find other models more suitable throughout the process.

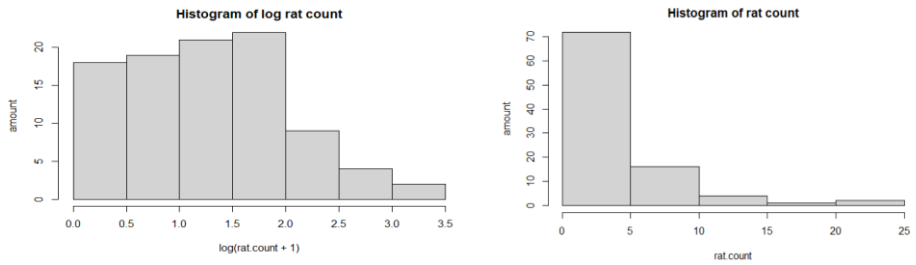
Histograms of the continuous variables:





- Possible outliers
- Most pixels are close in distance to these variables.

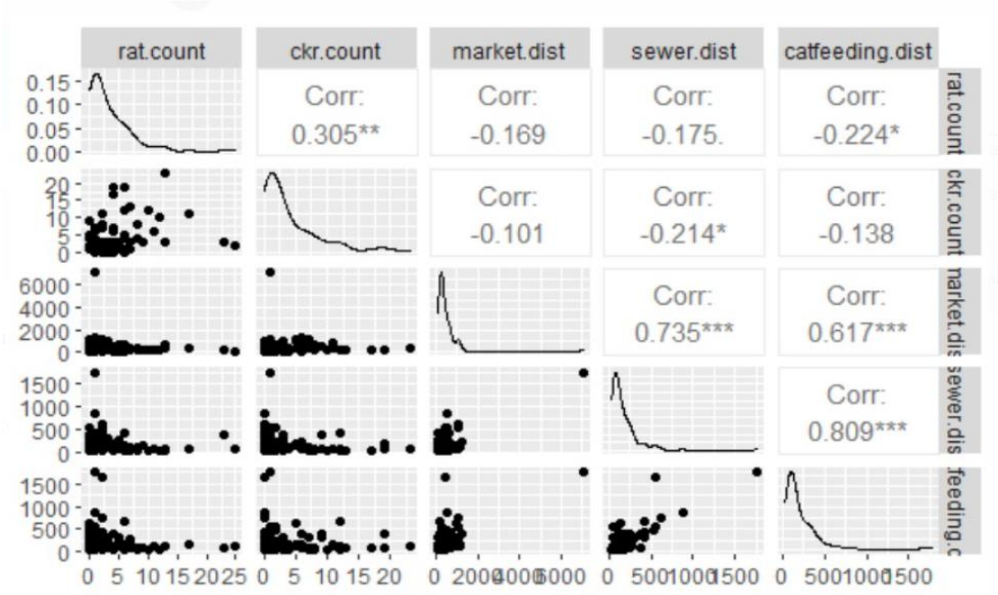
Histogram of rat count and log rat.count:



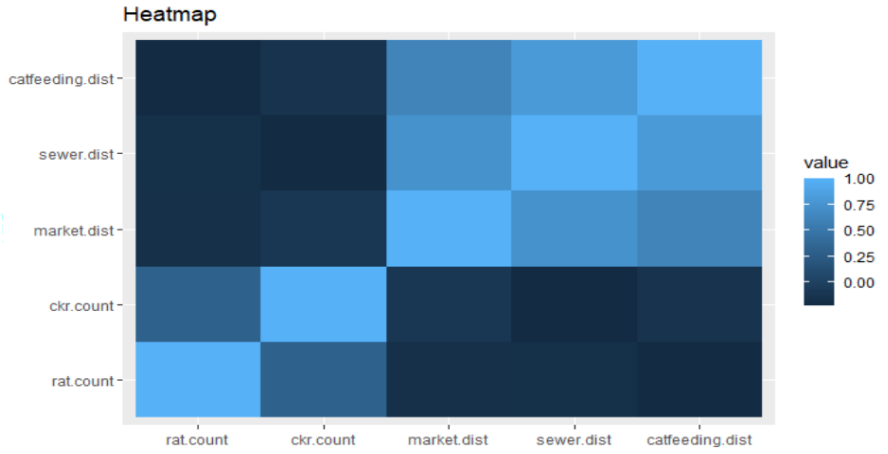
- Non-normal distribution

Pairwise scatterplots & correlations:

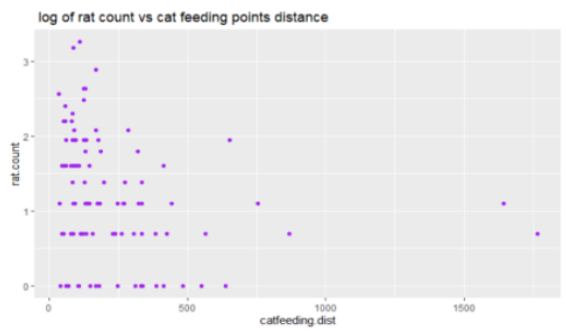
Note that all variables are continuous

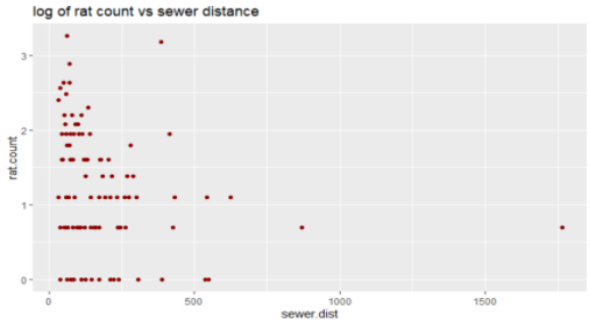


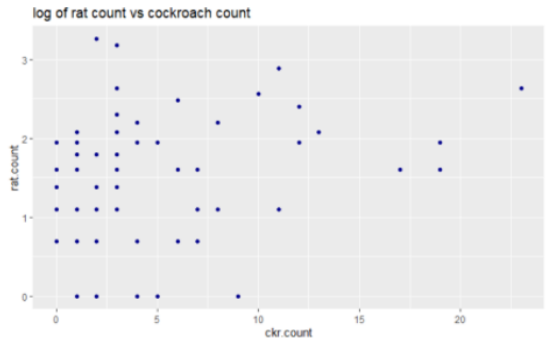
- Partially correlation between Rats and Cockroaches.
- Partially weak negative correlation between Rats and Cat feeding points.
- High correlation between variables may indicate multicollinearity (sewer.dist & catfeeding.dist).

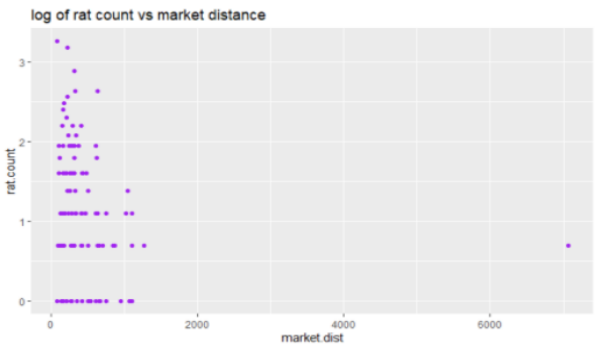


Now, we can observe relevant scatter plots in a clearer way to get a more genuine idea of the behavior our data through our explanatory variable's vs our explained variable.









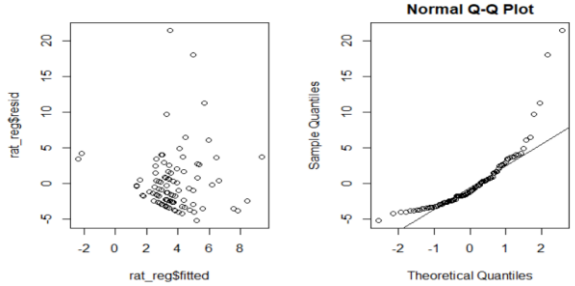
- Again, possible outliers in our data, we will get to that later.
- Some connection between the existence of rats and cockroaches.
- Partially weak negative correlation between Rats and Cat feeding points.

Regression

First, we should fit a linear regression model for our data (OLS) to observe the behavior of the data in this kind of model fitting and show that our initial thought of model fitting was correct, that is, to justify our GLM fit to this data.

```
Call:
lm(formula = rat.count ~ ., data = rat_data1)

Residuals:
   Min     1Q Median     3Q    Max
-5.193 -2.680 -1.173  1.454 21.517

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.2111955  0.7512826   4.274 4.76e-05 ***
ckr.count       0.2873140  0.0979081   2.935  0.00424 **
market.dist    -0.0006835  0.0008961  -0.763  0.44763
sewer.dist      0.0037213  0.0040730   0.914  0.36333
catfeeding.dist -0.0043276  0.0027761  -1.559  0.12253
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.312 on 90 degrees of freedom
Multiple R-squared:  0.1356,    Adjusted R-squared:  0.09719
F-statistic:  3.53 on 4 and 90 DF,  p-value: 0.01008
```

diagnostic plots (resid vs pred, npp of resids):



Interpretation:

The null hypothesis: $\hat{\beta}i = 0$, for each variable i. Therefore, we cannot reject the null hypothesis for all variables excluding the cockroach count variable. Moreover, the R-squared here is very small, indicating a weak goodness of fit in our linear model. Furthermore, we can see in the additional diagnostic plots a bad fit for this model (non-normal distribution).

Now, let us observe the linear model of a log-transformation of our explained variable. Here, we would take the log of the rat count and add 1 inside the log (to deal with pixels with 0 rat counts). We use this method to make sure that the right model to fit here would not be OLS.

```
Call:
lm(formula = rat.count ~ ., data = log_rat_data1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.46737 -0.46760 -0.07202  0.56996  2.10052

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.0995200  0.1362814   8.068 2.95e-12 ***
ckr.count        0.0570144  0.0177604   3.210  0.00184 **
market.dist     -0.0001205  0.0001626  -0.741  0.46060
sewer.dist       0.0004842  0.0007388   0.655  0.51394
catfeeding.dist -0.0006910  0.0005036  -1.372  0.17341
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7822 on 90 degrees of freedom
Multiple R-squared:  0.154,     Adjusted R-squared:  0.1164
F-statistic: 4.097 on 4 and 90 DF,  p-value: 0.004267
```

Similarly, this is not a good model to fit, the null hypothesis is not rejected, and we have a bad model to predict our data.
In conclusion, the linear model does not fit out data in the right way. In this case, we should find a proper GLIM model that fits our data.

We will first try to fit a Poisson model to our data.
The reason is that we are analyzing a count data and Poisson regression analysis is one of the ways to model these kinds of data. The definition of a Poisson distribution is "*a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space*". Note that the explained variable we are trying to predict here is the number of rats in a specified pixel, which fits this definition.
Note that, under the assumption of a Poisson model, the mean is equal to the variance (and both equal to the rate response).
In addition, in Poisson Regression, we have:

$$\mathrm{E}(Y \mid \mathbf{x}) = e^{\boldsymbol{\theta}' \mathbf{x}}$$

so,

$$\log(\mathrm{E}(Y \mid \mathbf{x})) = \boldsymbol{\theta}' \mathbf{x}$$

and the summary of the model fitted to our data will be:

```
glm(formula = rat.count ~ ., family = poisson(link = log), data = rat_data1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0885  -1.5462  -0.5401   0.6675   6.7019

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.6258919  0.1361237  11.944  < 2e-16 ***
ckr.count        0.0534658  0.0094618   5.651  1.6e-08 ***
market.dist     -0.0009025  0.0002539  -3.555 0.000379 ***
sewer.dist       0.0006790  0.0006690   1.015 0.310070
catfeeding.dist -0.0018192  0.0005527  -3.291 0.000998 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 416.57  on 94  degrees of freedom
Residual deviance: 326.97  on 90  degrees of freedom
AIC: 570.96

Number of Fisher Scoring iterations: 6
```

Here, we can start getting closer to a fine fit of a model. Most of the P-values are small enough as desired and we can start and think of rejecting the null hypothesis. Although, we should first check for overdispersion in our model.

Checking for overdispersion by estimating φ:

In a Poisson model, we expect that the mean & variance values to be equal to the rate value and therefore we expect the residual deviance to be equal (or close to) our degrees of freedom. For estimation we check the square of the Pearson's residuals divided by the degrees of freedom of our Poisson model.

```
> disp <- (sum(residuals(rat_reg1, type = "pearson")^2))/90
> disp
[1] 13.14652
```

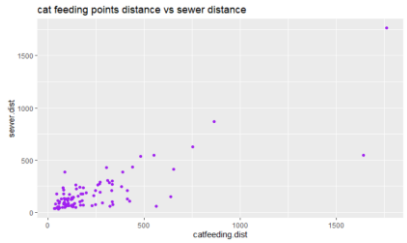
The dispersion parameter ("disp "), with 90 degrees of freedom in our model, should be equal (or close to) 1.
In conclusion, we have found overdispersion in our model (the value estimation of phi is much larger than 1).
In these cases, we should check the "Quasi-Likelihood" model, in our case we would check the "Quasi-Poisson" model and find our phi estimator.

So, for our new quasi-Poisson model we have:

```
glm(formula = rat.count ~ ., family = quasipoisson(link = log),
    data = rat_data1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0885  -1.5462  -0.5401   0.6675   6.7019

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.6258919  0.4935595   3.294  0.00141 **
ckr.count      0.0534658  0.0343066   1.558  0.12263
market.dist   -0.0009025  0.0009206  -0.980  0.32954
sewer.dist     0.0006790  0.0024255   0.280  0.78015
catfeeding.dist -0.0018192  0.0020042  -0.908  0.36646
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 13.14654)

    Null deviance: 416.57  on 94  degrees of freedom
Residual deviance: 326.97  on 90  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
```

- We have found the estimator : $\hat{\varphi} = 13.14$.

- We can see that the sewer.dist has a large p-value, the reason may be the existence of outliers in our data. Another reason can be the existence of multicollinearity between variables in our model. In GLIM the VIF and TOL tests for multicollinearity do not hold so we can use high correlation as an indication of multicollinearity. If we go back to our correlation graph we can see correlation of more than 0.8 between catfeeding.dist and sewer.dist.



  The solution here would be fitting a model, discarding the catfeeding.dist variable (we will get to this later).

- Another possible explanation is the existence of interaction between variables. A solution to this problem may be to fit a model with an interaction variable. After trying several other possible interaction variables, I found that the best interaction variable to fit a model to our data is $\sqrt{\text{catfeeding.dist} + \text{sewer.dist}}$.

Note that I would present conclusions to both models (one with the interaction variable and one excluding the cat feeding variable).

The model with the interaction variable:

```
glm(formula = rat.count ~ ckr.count + market.dist + sqrt(sewer.dist +
    catfeeding.dist), family = quasi(link = "log", variance = "mu"),
    data = rat_data1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2811  -1.6495  -0.4885   0.6335   7.0584

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      2.0861492  0.6847684   3.047  0.00303 **
ckr.count                        0.0460671  0.0337130   1.366  0.17516
market.dist                     -0.0009402  0.0009222  -1.019  0.31069
sqrt(sewer.dist + catfeeding.dist) -0.0374269  0.0358903  -1.043  0.29980
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasi family taken to be 13.42995)

    Null deviance: 416.57  on 94  degrees of freedom
Residual deviance: 328.10  on 91  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
```

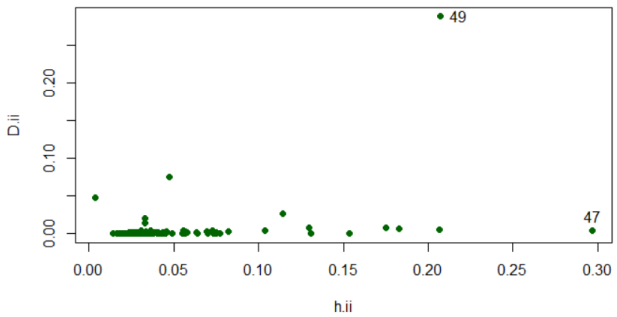
We can see that we have improved the p-values. We might come back to this model after finding (and removing) relevant outliers if our previous (original) quasi-model fails to reject the null hypothesis (spoiler alert: it will fail).

Outliers

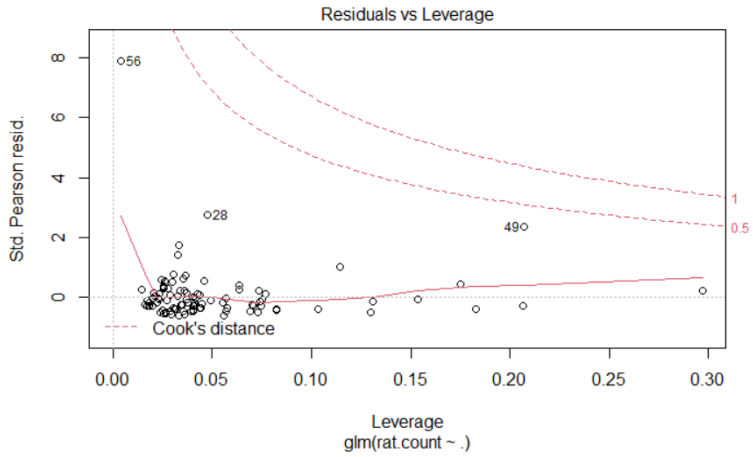
Now, we would check for outliers in our data. We are searching for observations with extreme behavior compared to the other observations for several possible reasons and try to explain and analyze the causes.
Note that, we are seeking for observations that may be problematic for our purpose. These possible observations, if exists, can cause damage to are interpretation because they fail to be decent representatives of our data. On the one hand, removing some of these observations from our data can lead to ***overfitting*** problems and on the other hand, leaving them in our data may interfere with the reliability of our conclusions. So, we need to be careful with what we remove from our data, taking in consideration that with additional data we might act differently.

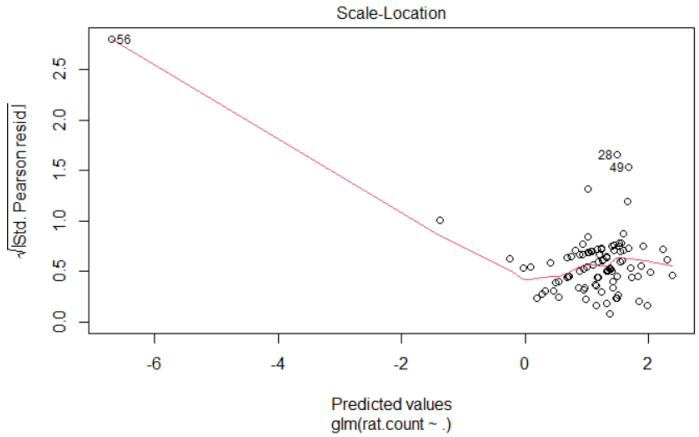
For this, we first observe the Leverage and the Cook's distance for each observation. Then we can add another dimension by adding the Std.Pearson's residuals. The third graph is the Pearson residuals over the predicted values. Finally, we can observe the data related to each of the different variables.

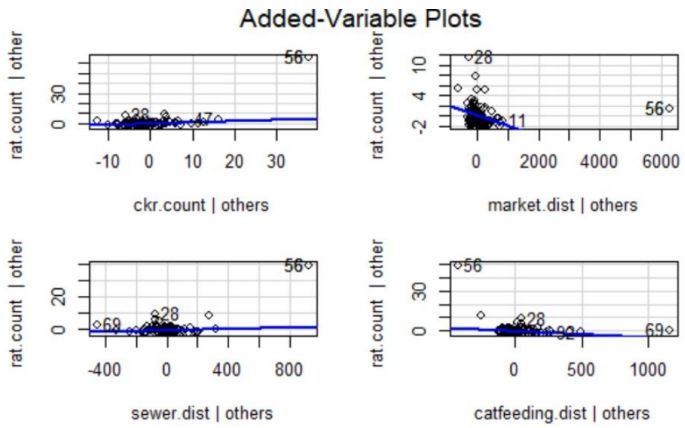
```
D.ii <- cooks.distance(quasi_rat_reg)
h.ii <- hatvalues(quasi_rat_reg)
```



Here, we can see that observation number 47 has a high leverage, the possible reason is the large number of cockroaches found there. We can also see that observation number 49 has extreme value of Cook's distance.







In conclusion, we can now identify possible outliers:

We can see that observations 28 and 49 have relatively large residuals, although their predicted values fall where most of the other predicted values are. The reason is they are the observation with the most rats counted (much more than we expect to find there).

We can also see that observation 56 has a very large residual and an extreme predicted value. It is important to try to get a better understanding of the reasons for the extreme results.

	rat.count	ckr.count	market.dist	sewer.dist	catfeeding.dist
56	1	1	7058.24138	1764.12644	1764.12644

It seems that this observation has extreme distances from the nearest market, cat feeding point and sewer.

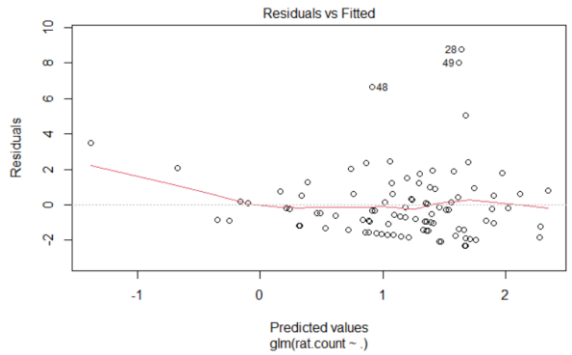
- We can try to check if there had been a mistake in the data reporting, for now we can assume that there is not.

In result, we should fit the model with and without the outliers and compare the results. Note that removing observation has a cost of losing possible data or causing overfitting problems.
We can compare the different fitted models (excluding different possible outliers).
Note that, in this case we can't use the AIC test (due to our quasi-likelihood model) to compare the different models, so we would compare the $\hat{\beta}$i regarding their values and significance.

Fitting a model without observation number 56:

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.8213849  0.2796519   6.513 4.26e-09 ***
ckr.count       0.0501002  0.0191955   2.610   0.0106 *
market.dist    -0.0013405  0.0005567  -2.408   0.0181 *
sewer.dist      0.0002530  0.0013913   0.182   0.8561
catfeeding.dist -0.0016919  0.0011323  -1.494   0.1387
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 4.117296)
```

The residuals vs the predicted values without observation number 56:



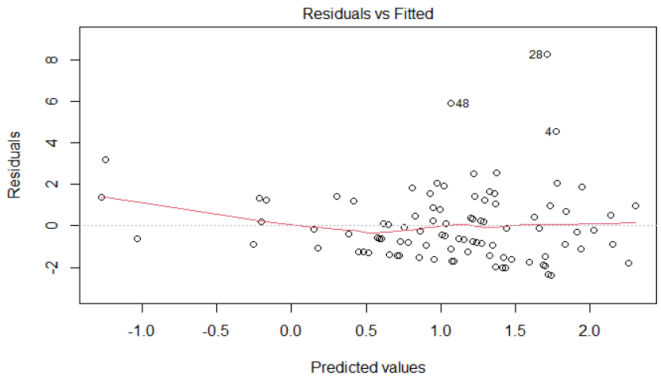
We can see that the model looks better now due to the fact that the relevant p-values are smaller (except for the sewer.dist - we will deal with it later) , so we can check if fitting this model without observations 28 and 49 would improve our model, in addition of removing observation number 47 which has a high leverage (more than 2*sum(h.ii)/95).
Fitting the model without observations 28,47,49,56:

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.6354863  0.2415608   6.770  1.5e-09 ***
ckr.count       0.0508494  0.0180696   2.814  0.00606 **
market.dist    -0.0009391  0.0004288  -2.190  0.03122 *
sewer.dist     -0.0021408  0.0012708  -1.685  0.09569 .
catfeeding.dist -0.0002893  0.0008185  -0.353  0.72466
---
```

Now our model seems to be even worse (and with less data). We have several options as mentioned above; we can now try to fit a model discarding the cat feeding distance variable or try fitting the model presented above with the relevant interaction variable.

1. Regarding the first option, we can fit a model without observations 56 & 49:

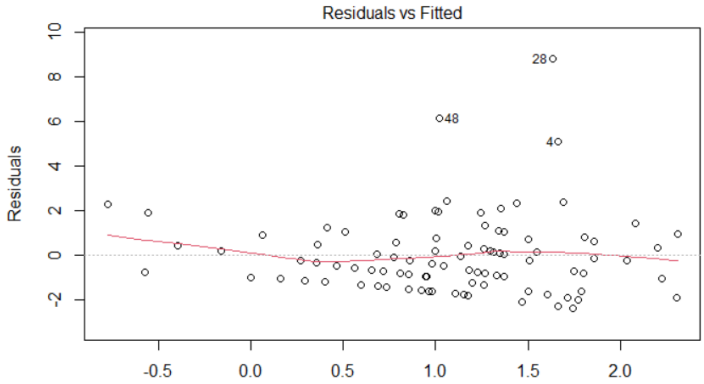
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.9112576  0.2502385   7.638 2.39e-11 ***
ckr.count    0.0444806  0.0166316   2.674  0.00891 **
market.dist -0.0013085  0.0004736  -2.763  0.00696 **
sewer.dist  -0.0029004  0.0011092  -2.615  0.01049 *
---
```



We can see that the null hypothesis is rejected, we got most of our data (rejected only 2 extreme observations), and very small p-values. On the other hand, we lost information regarding cat feeding points.

2. Regarding the second option, we can fit a model without observations 56 & 49:

```
Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     2.2948598  0.3529610   6.502 4.48e-09 ***
ckr.count                       0.0477592  0.0167389   2.853  0.00538 **
market.dist                    -0.0011608  0.0004929  -2.355  0.02073 *
sqrt(sewer.dist + catfeeding.dist) -0.0505900  0.0192309  -2.631  0.01004 *
```



Here we can see that the null hypothesis is rejected, we got most of our data (rejected only 2 observations), and very small p-values. On the other hand, we are using an interaction variable and not using the actual data we have for each variable.

- Note that I chose to exclude these observations due to the extreme values of observation 56 and due to the extreme Cook's distance of observation 49. In addition, I compared all the possible combinations of observations excluded (amongst identified outliers) in each model presented and got the model with the best results when these observations had been removed.

I would prefer the model including the interaction variable. It seems like we lose minimum information and obtain decent results.

Despite that, I used the γ test method used in class to check irregularity and did not get expected results.

For observation 56, under the interaction model, we get $\gamma = 0.001$ and for observation 49 we get $\gamma = 0.002$.

Finally, we can check again for our dispersion parameter and find our new φ estimator for the final model:

```
> disp <- (sum(residuals(quasi_rat_reg_clean, type = "pearson")^2))/89
> disp
[1] 3.192687
```

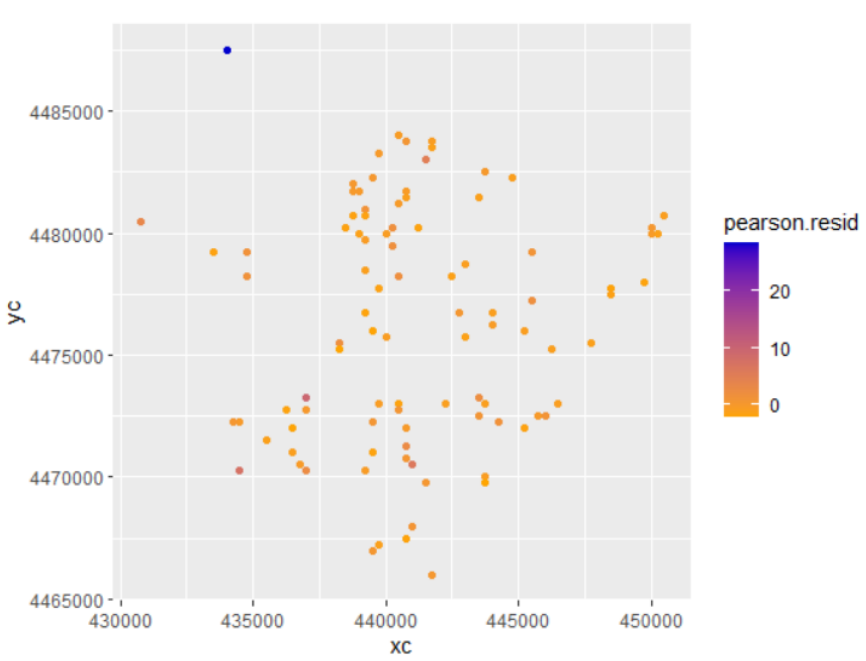
We have $\hat{\varphi} =$

To the fitted model:

```
glm(formula = rat.count ~ ckr.count + market.dist + sqrt(sewer.dist +
    catfeeding.dist), family = quasi(link = log, variance = mu),
    data = rat_data_clean)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3833  -1.4934  -0.2797   0.7182   6.2844

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                        2.2948598  0.3529610   6.502 4.48e-09 ***
ckr.count                          0.0477592  0.0167389   2.853  0.00538 **
market.dist                       -0.0011608  0.0004929  -2.355  0.02073 *
sqrt(sewer.dist + catfeeding.dist) -0.0505900  0.0192309  -2.631  0.01004 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasi family taken to be 3.192707)

    Null deviance: 367.48  on 92  degrees of freedom
Residual deviance: 262.45  on 89  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

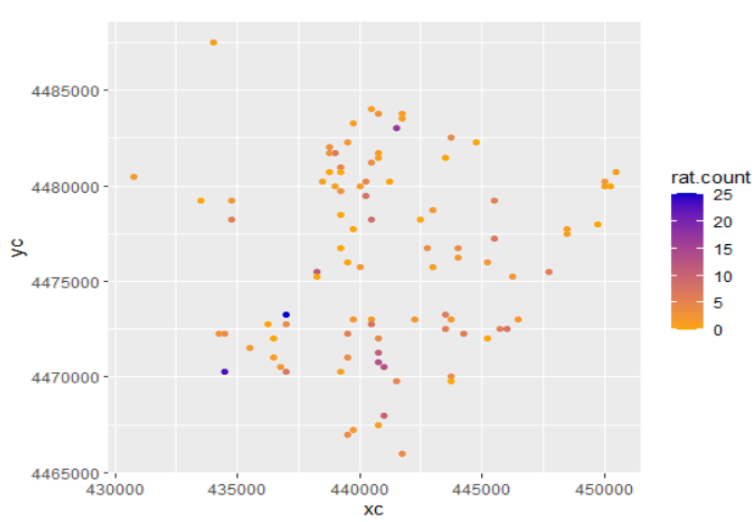
Spatial analysis

Now, after we have a good understanding of the data, we may think of additional variables which may improve our model fitting. In our original data we had geographic coordinates of the exact location of each pixel. We may think that there might be some correlation between the existence of rat and the area in Madrid. For example, we my think that it is possible that the more north you travel in the city, the more rats you are expected to find. Furthermore, it would be reasonable to assume that more rats would be counted in poor neighborhoods with worse living conditions, due to dilapidated infrastructure, which may be localized in specific areas.
If we find such connections we can think of adding these x,y coordination's to our data.
We can look at the graph of the Pearson residuals vs our coordinates and check if there is a geographic connection to some of the unique pixels. Here, we might also find reasonable explanations for some of the possible outliers we found.



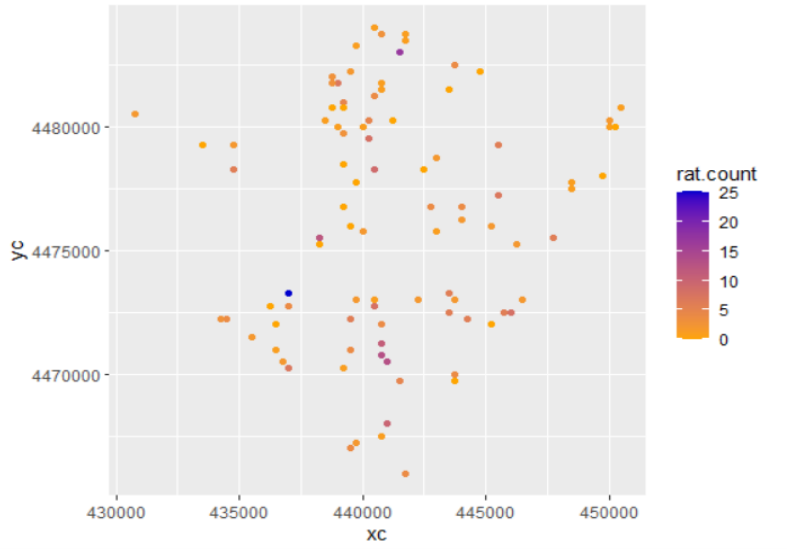
- We can see that observation 56 is in fact distant from the other observed pixels, that may explain why it is very far from a market/cat feeding point/sewer.
- Lack of unique geographic trend.
- It seems that there is no geographic connection in this data but remember that we fit models discarding outliers.

Now, we can check for some interesting behavior of the number of rats geographically.



- Seems to lack some unique significant trend, even though in the south we expect to find slightly more rats.

We can observe the data without the observations we excluded:



Conclusion:

Because there is no spatial dependency or unique trend in these cases, we will not add x&y coordinates to our model.

Final conclusions

In this paper we analyzed data of 95 observations of rat sightings in Madrid and tried to find a proper relation between the counted rats in a pixel to the various information we got about that pixel. Our goal was to be able to predict the number of rats expected to be found in a pixel given some information regarding that pixel.
After observing preliminary information extracted from the original data, we chose the variables we thought would make sense with predicting our data and started to fit regression models to our data.
First, despite having a "Poisson related hunch" we tried to fit a linear model (using OLS) to our data and concluded that it was a bad fit to our data, due to high p-values for our covariates and extremely low R-squared values. Then, we falsely assumed that maybe a simple log transformation over our explained variable might do the job, but the results were poor again.
Next, we tried fitting a model using Poisson regression. In this case, we started to see a reasonable possible fitting direction although not as good as expected due to the existence of "overdispersion" in our model.
As a result of having overdispersion in our Poisson model, we used the Quasi-Likelihood method to overcome this problem and found our estimation of phi. Although we no longer had overdispersion in our model, we lacked the significance we expected to obtain in our model and failed to reject the null hypothesis.
We assumed that our problem achieving a decent model to fit our data can be a cause of existence of outliers or the existence of interaction between variables in our model (or both). We first checked for possible outliers in our data using several methods, for example observing the Cook's distance, leverage and residuals of each observation. During this process, we found several observations with different extreme values and tried to find the best fit of a model that rejects some combination of these observations but does not fall to overfitting. We found out that even though our model improves significantly when we exclude two extreme observations, there remains a trade-off between two highly correlated variables.
Due to these finding we presented two models (in addition to the model without excluding outliers).
The first model, under the assumption that it is possible that the nonsignificant variable (cat feeding point) is not important or dependent with another variable (sewer.dist), we excluded it from our model and got a decent significant model which fits our data.
The second model, under the assumption that we do not want to lose information that seems important in addition to the possibility that there might be interaction between two variables, contains an interaction variable instead of the highly correlated variables.
Then we found our new estimation of φ and presented our dispersion parameter.
Finally, we justified not including geographic variables in our model by a spatial analysis. We helped explain the reasons of the extreme values of some of the outliers and saw there is no significant special trend in the existence of rat in Madrid (geographically).
In my opinion, the second model, with the interaction variable, is the best fit to our data. We used almost all the data which seemed relevant for predicting the number of rats in a pixel, excluding only two extreme outliers who fail to represent our data decently.