Regression Analysis of Crime Rates Based on Alcohol Establishments, School Attendance, and Worship Participation Factors

Emma Boehly, Constance de Trogoff, Sander Miesen, Gaëlle Verdon

2024-07-03

Introduction

Investigating the causes of crime is a complicated task. Indeed, many external factors, such as civil wars and worker strikes could influence any region's crime rates drastically, and these factors are hard to control for. Our dataset, collected by John Clay between 1849 and 1853, has the particularity of having very few of those disturbances as there were, in his own words, "no political or social excitement [and] no cessation of the employment". Additionally, in 1851 there was a Census, thus making the recording of the number of inhabitants as accurate as possible. These two factors make this an ideal case study, controlling for many external variables.

In this report, we aim to investigate how counties' number of criminals per 100'000 inhabitants are affected by the number of ale and beer houses per 100'000 inhabitants, as well as the number of people attending at school per 10'000 inhabitants and the number of people attending at public worship per 2'000 inhabitants. These variables are interesting as they relate to three more general socio-cultural themes which often come up when talking about crime, namely drunkenness, education and religion. Through the running of regression models, we hope to find significant relationships between the variables and the crime rate, which could eventually help set up measures to decrease it.

EDA

To start our data exploration, we first performed a univariate data analysis of the predictors and of the response variable of interest, the number of criminals per 100k per county. We have 3 numerical features : the number of ale and beer houses per 100k per county, the number of public school attendants per 10k per county, and the number of public worship attendants per 2k per county, and 3 categorical features : the county names, the region names and the region codes. By looking closely ar the data, we noticed that each sample corresponded to a different county name and thus we ignored this last feature for the rest of our analysis.

We also remarked that each region name corresponded to a region code, except for South Midland which had the same region code as South Eastern, thus we replaced the region_code “1” of South Eastern to factor “0”. Occurrences of categorical features are indicated in Table 1.

Regarding the three numerical predictors, we first computed metrics which informed us on the center of their distribution (the mean and the median), and its spread (with the minimum and maximum value, the lower percentile 25 and the upper percentile 75 as well as the standard deviations), all of which are displayed in Table 2. While the mean is very sensitive to outliers, the median is not, and thus by comparing the two we get insights into the presence of outliers. We can observe that the mean and the median are quite similar relative to the spread for all three numerical features, hinting at the absence of outliers. In order to have a visual summary of the distributions, we made boxplots of the predictors, shown in Figure 1. They appear to be quite symmetric, although they are slightly left-skewed for the ale and beer houses and the worship attendants, and slightly right-skewed for the public school attendants.

Region name	Occurence	Region code	Occurence
South Midland	7	1	7
West Midland	7	4	7
South Eastern	6	0	6
North Midland	5	3	5
South Western	5	5	5
Northern	4	8	4
Eastern	3	2	3
North Western	2	6	2
York	1	7	1

Table 1: Occurrence of the two categorical features of interest : Region name and Region code

	Ale/Beer houses	School Attendants	Worship Attendants
Min	87.0	560.0	434.0
1st Qu.	209.0	880.0	654.5
Median	407.0	965.0	801.0
Mean	374.9	957.8	780.1
3rd Qu.	490.8	1082.5	912.0
Max	708.0	1250.0	1136.0
SD	165.0	161.4	172.5

Table 2 : Numerical univariate analysis of the three predictors : Number of Ale/Beer houses per 100k, Public school attendants per 10k and Public worship attendants per 2k

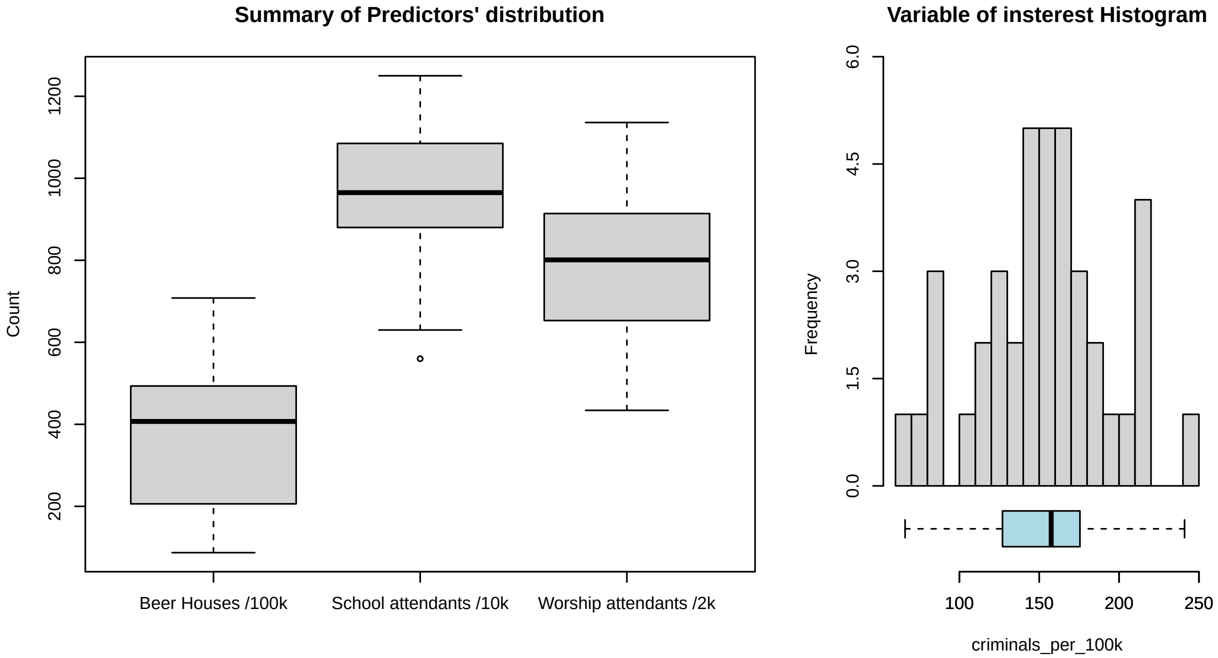

Figure 1 : Univariate graphical analysis of numerical variables. On the left, boxplots of the three predictors : Number of Ale/Beer houses per 100k, public school attendants per 10k and public worship attendants per 2k. On the right, histogram and boxplot of the variable of interest : the number of criminals per 100k

Figure 1 : Univariate graphical analysis of numerical variables. On the left, boxplots of the three predictors : Number of Ale/Beer houses per 100k, Public school attendants per 10k and Public worship attendants pers 2k. On the right, histogram and boxplot of the variable of interest : the number of criminals per 100k

Next, to get a better look at the response variable of interest, i.e. the number of criminals per county, we visualized its distribution with a histogram and a boxplot. As we can see in Figure 2, it is slightly left skewed. For a first naive comparison of its distribution across regions, we displayed separate boxplots for each region in Figure 3. We can observe clear differences between the number of criminals' medians, and thus we ordered them in ascending order. To get an initial intuition about how the predictor variables will influence the number of criminals, we looked at the counties with respectively the lowest (region 8, Northern, with a value of under 100) and highest (region 4, West Midland, with a value of over 200) median number of criminals. In Figure 3, we can see that the distributions showed a net higher number of ale and beer Houses and a higher number of worship attendants in region 4 compared to region 8, so we can naively hypothesize that these two predictors might be positively correlated with the number of criminals. The number of attendants at public schools seem to be the same for both however, so there is nothing we can say about that yet.

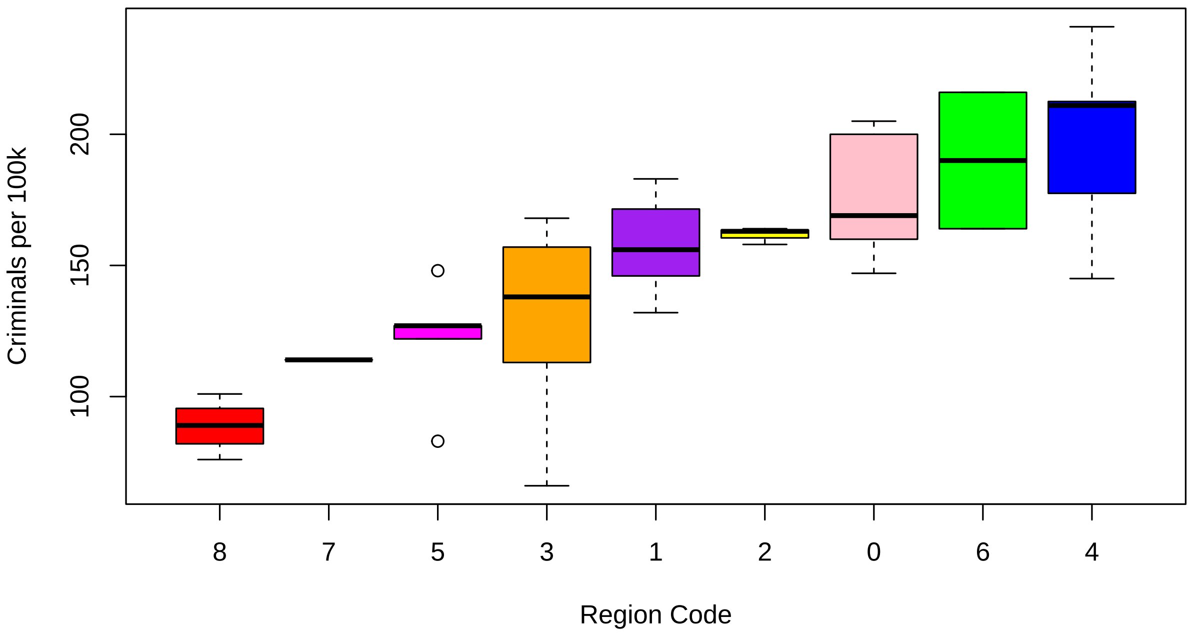

Figure 2 : Number of criminals per 100k by region code. Colors of the region code remain the same for the whole analysis

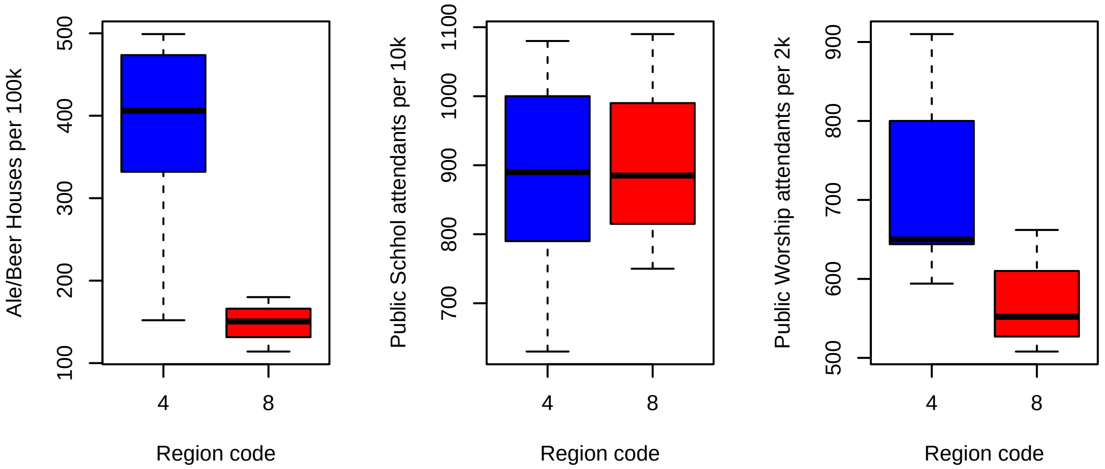

Figure 3: Number of ale/beer Houses per 100k, public School attendants per 10k and public Worship attendants per 2k (the three predictors) for region Codes 4 and 8.

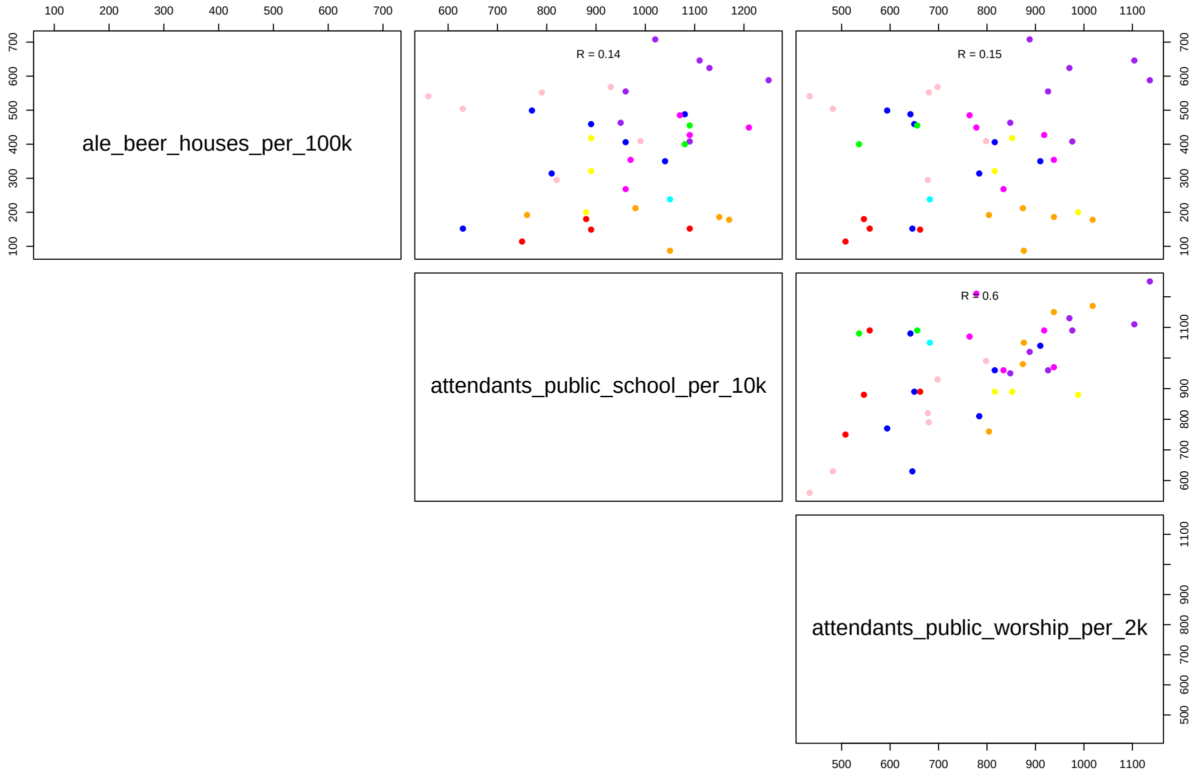

Figure 4: Pairwise correlation plot between the three predictors : Ale/Beer Houses per 100k, public school attendants per 10k and public worship attendants per 2k. Pearson correlation coefficients are computed independently of the region code. Colors correspond to region codes as in Figure 2.

To better visualize how the predictors are related with each other, we presented them in a pairwise correlation plot along with their pearson correlation coefficients which were computed independently of the region code, and this is shown in Figure 4. We can notice a positive correlation between the number of public school attendants and the number of public worship attendants which may indicate the necessity of removing one of the two in the following linear models in order to have a good balance between model accuracy and model simplicity.

Model Fitting

In this project, we aim at finding a linear regression model predicting the number of criminals per 100'000 people. As stated in the project description, we will only consider numerical variables in our model, so only the number of ale and beer Houses per 100'000 people, the number of attendants at public school per 10'000 people, and the number of attendants at public worship per 2'000 people. In the following, we will first look at the model prediction based on these 3 variables and then we will try to find the best model possible by following a backward model selection process. This process will be based on adjusted R squared method

(R^2_{adj} in the following) and the Akaike's Information Criterion (AIC in the following). We will try to lowering the AIC while increasing the adjusted R-squared of our models.

In this project, we aim at finding a linear regression model predicting the number of criminals per 100 000 people. As stated in the project description, we will only consider numerical variable in our model, that is to say, only Ale/Beer Houses per 100 000 people, Attendants at public school per 10 000 people, and Attendants at public worship per 2000 people will be taken into account in our model. In the following, we will first look at the model prediction based on these 3 variables and then we will try to find the best model possible by following a backward model selection process. This process will be based on adjusted R squared method (R^2_{adj} in the following) and the Akaike's Information Criterion (AIC in the following). We will try to lowering the AIC while increasing R squared adjusted in our models.

Model 1:
$Y = 172.88 + 0.12\beta_0 + 0.10\beta_1 + 0.039\beta_2$ This model gives a R^2_{adj} of 0.2619 and an AIC of 405.0494. Now that we have out reference model, let's try to optimize it. In the exploratory data analysis, we saw that the number of attendants at public school per 10k and the number of attendants at public worship per 2k have a positive correlation of 0.6. Therefore, we decided to remove the latter in the next model.

Table 3 : Summary of variable names

Model 2:
$Y = 178.81 + 0.13\beta_0 - 0.077\beta_1$
This model gives a R^2_{adj} of 0.26 and an AIC of 404.041. We can see that the model has a highest R^2_{adj} and lower AIC, which means that the model is better than model 1. In the next model, we then wanted to see whether the number of attendants at public worship per 2k contributes more to the performance of the model compared to the number of attendants at public school per 10k.

Model 2:
$Y = 178.81 + 0.13\beta_0 - 0.077\beta_1$
This model gives a R^2_{adj} of 0.26 and an AIC of 404.041 . We can see that the model has a highest R^2_{adj} and lower AIC, which means that the model is better than model 1. In the next model, we then wanted to see whether attendants at public worship per 2k contributes more to the performance of the model compared to attendants at public school per 10k.

Model 3:
$Y = 121.26 + 0.12\beta_0 - 0.017\beta_2$ This model gives a R^2_{adj} of 0.177 and an AIC of 408.52 which correspond to poorer results compared to model 2. To make sure that the number of attendants at public school per 10k really contributes to the improvement of the model, we decided to remove it in the next model. The number of attendants at public worship per 2k is also removed in the next model model since we already demonstrated that it wasn't contributing to the increase in performance.

Model 4:
$Y = 109.34 + 0.12\beta_0$
This model gives a R^2_{adj} of 0.19 and an AIC of 406.75 . Since both the R^2_{adj} and the AIC are respectively lower and higher than for model 2, the number of attendants at public school

per 10k does contribute to the increase of performance of our model. Finally, we wanted to make sure that the number of attendants at public school per 10k is not the only contributor to the performance of our model. To do that, we will only consider this variable in the last model.

Model 4:
$Y = 109.34 + 0.12\beta_0$
This model gives a R^2_{adj} of 0.19 and an AIC of 406.75 . Since both the R^2_{adj} and the AIC are respectively lower and higher than for model 2, it means that attendants at public school per 10k does contribute to the increase of performance of our model. Finally, we wanted to make sure that attendants at public school per 10k is not the only contributor to the performance of our model. To do that, we will only consider this variable in the last model.

Model 5:
$Y = 209.42 - 0.059\beta_1$
This model gives a R^2_{adj} of 0.15 and an AIC of 414.22 . Since both the R^2_{adj} and the AIC are respectively lower and higher than for model 2, it means that the number of ale and beer houses per 100k also contributes to the performance of our model.

Model	R^2_{adj}	AIC
Model 1	0.2619	405.0494
Model 2	0.2638	404.0409
Model 3	0.1767	408.5154
Model 4	0.1936	406.7524
Model 5	0.1533	414.2217

Table 4 : Summary of the R^2_{adj} and AIC for each model considered

Model selection and Model Assessment

In the last section, we computed different models to determine which one is better. We saw that model 1 and model 2 gave us very close results when looking at the adjusted R-squared. However, when looking at AIC values, we can see that model 2 has a lower one and we therefore decided to consider model 2 as our best model. The final model that we selected is $Y = 178.81 + 0.13\beta_0 - 0.077\beta_1$ (model2)

However before accepting our model, it is essential to verify that the model's assumptions hold. These include uncorrelated, homeoscedastic, normally distributed erros with mean 0. By looking at the diagnostic plots, especially the residuals vs fitted plot, we can observe that the mean of the residuals is slightly below 0 and not perfectly homeoscedastic with 3 points (23, 24, 34) having a high residual value. Furthermore, the QQ-normal plot of residuals showed approximately normally distributed errors with a slight long tail on the right and three serious under-predictions corresponding to the previously mentionned points. Colors of the different regions were displayed on the plots in order to see if those specific points belong to a particular region. It appeared that most of the points of the right long tail including points 23 and 24 belong to the region of West Midland (region code 4) which has a particular high number of criminals (see EDA). In the same way, point 34 belongs to the region of North Western (region code 6) which is the second region with the highest number of residuals. As a result, we could hypothesize that the peculiar number of criminals in those two regions might not fully explained by our model. Other diagnostic plots on the presence of influential points were also visualized although not displayed here and no specific outliers were found although point 24 of West Midland appeared to be the most influential one.

However before accepting our model, it is essential to verify that model assumptions hold. These include uncorrelated, homeoscedastic, normally distributed erros with mean 0. By looking at the diagnostic plots, especially the residuals vs fitted plot, we could observe that the mean of the residuals is slightly below 0 and not perfectly homeoscedastic with 3 points having a high residual value : Points 23, 24 and 34. Furthermore, the QQ-normal plot of residuals showed approximately normally distributed errors with a slight long tail on the right and three serious under-predictions corresponding to the previously mentionned points. Colors of the different regions were displayed on the plots in order to see if those specific points belong to a particular region. It appeared that most of the points of the right long tail including points 23 and 24 belong to the region of West Midland (region code 4) which has a particular high number of criminals (see EDA). In the same way, point 34 belongs to the region of North Western (region code 6) which is the second region with the highest number of residuals. As a result, we could hypothesize that the peculiar number of criminals in those two regions might not fully explained by our model. Other diagnostic plots on the presence of influential points were also visualized although not displayed here and no specific outliers were found although point 24 of West Midland appeared to be the most influential one.

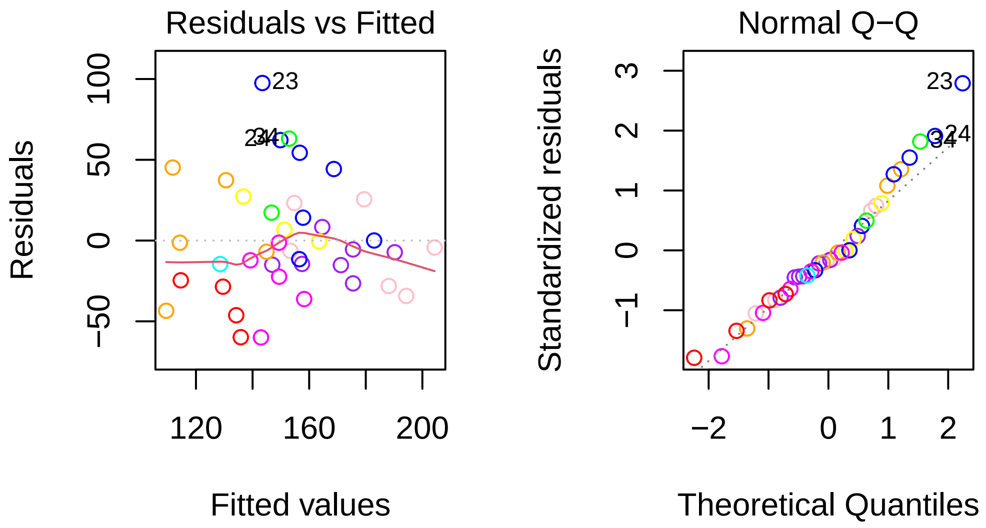

Figure 5: Diagnostic plots of model 2 $Y = 178.81 + 0.13\beta_0 - 0.077\beta_1$. On the left is the

Residuals vs Fitted plot. On the right, is the QQ-normal plot of residuals. Colors correspond to region codes as in Figure 2.

Conclusion

The goal of the project was to investigate some possible causes of crime in English counties in the 1850s, namely drunkenness, non-attendance to school and non-attendance to public worship.

Through exploration of the data, we saw that the number of criminals was clearly different from a county to another, which fueled our follow-up research.

Comparing the counties with highest and lowest crime rates, we saw that the number of ale and beer houses and the number of people attending public worship might be positively correlated to the number of criminals, and thus with crime rate, although the results were not significant, while the number of attendants to public school seemed to be the same.

Furthermore, we noticed that there was a strong positive correlation between the number of people attending public school and the number of people attending public worship, which helped guide our model selection. Indeed, in our final model, we only included the number of beer and ale houses and the number of people attending public school, as the number of people attending public worship did not seem to contribute much to the explanatory power of the model.

In the end, we found that the best model, chosen by considering both the explained variance (R^2_{adj}) and an information criterion (AIC), was $Y = 178.81 + 0.13\beta_0 - 0.077\beta_1$, where Y is the number of criminals per 100k, β_0 is the number of ale and beer houses per 100k and β_1 is the number of people attending public school per 10k. This final model indicates that the number of criminals seems to increase as the number of ale and beer houses increases, and decreases as the number of people attending public school increases. This first part is in accordance to the first hypothesis, while the second contradicts the second hypothesis that we made during exploratory data analysis. The reason for this is of course that our hypotheses were based on data from two particular counties only, while the final model considers all of them, reminding us of the importance of basing our generalized results on data from the whole population. Finally, we saw that the model assumptions were not perfectly met for some of the counties' data (West Midland, North Western), and that the model was therefore maybe not the best to explain the number of criminals in those counties.