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

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

Introduction

Investigating the causes of crime is a complicated task. 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 perform a univariate data analysis of the predictors and of the response variable of interest, the number of criminals per 100k per county. For the sake of simplicity, we will from now on refer to this variable as the number of criminals, and to the three predictors as the number of ale and beer houses, the number of public school attendants and the number of public worship attendants.

We have 3 numerical features : the number of ale and beer houses, the number of public school attendants and the number of public worship attendants, as well as 3 categorical features : the county names, the region names and the region codes. By looking closely at the data, we notice that each sample corresponds to a different county name and thus we can ignore this last feature for the rest of our analysis. We also see that each region name

corresponds to a region code, except for South Midland which has the same region code as South Eastern. Therefore, we replace the factor "1" of South Eastern to the factor "0". Occurrences of categorical features are indicated in Table 1.

Regarding the three numerical predictors, we first compute metrics which inform 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. In order to have a visual summary of the distributions and easily notice outliers, we can look at the boxplots of the predictors, shown in Figure 1. The distributions 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. We can also notice an outlier for the number of public school attendants, which we will need to watch out for in later analyses.

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

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

Next, to get a better look at the response variable of interest we visualize 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 display separate boxplots for each region in Figure 3, ordering them in ascending order of the distributions' medians. We can observe clear differences depending on the region : the region with code 8 has half the number of criminals than the region with code 4. To get an initial intuition about how the predictor variables will influence the number of criminals, we look 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 Fig.3 (located in the Appendix), we can see that the distributions show 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.

	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

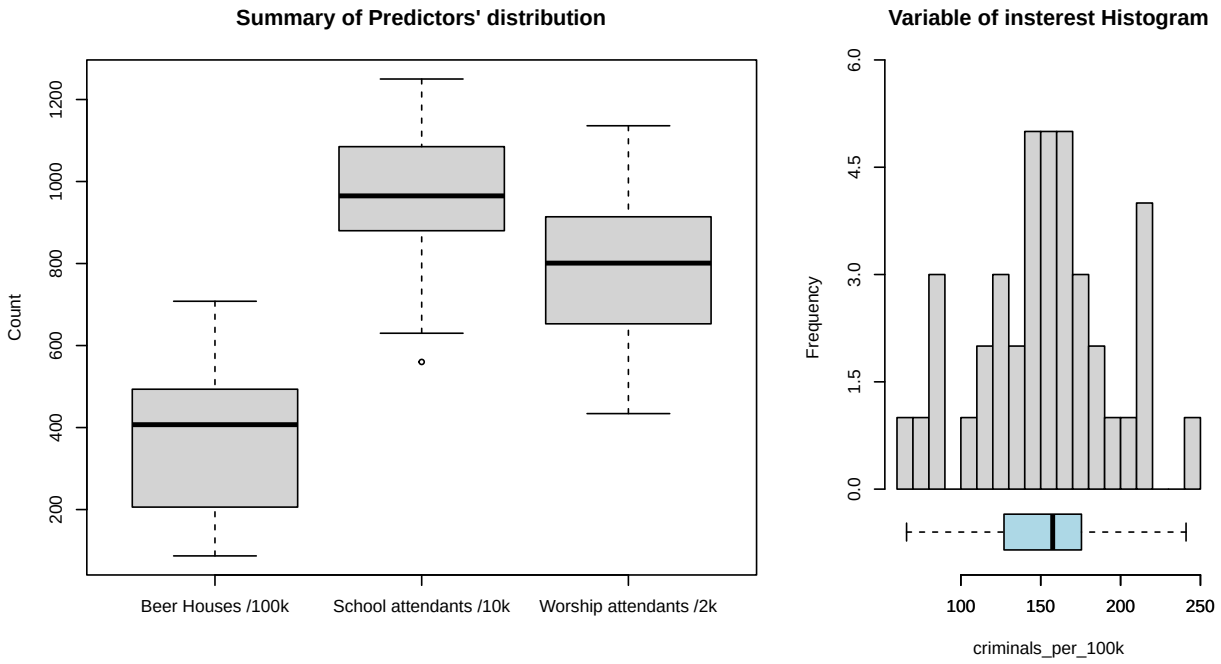

Figure 1 : Univariate graphical analysis of numerical variables. On the left, boxplots of the three predictors : Number of Ale/Beer houses per 100k, public school attendants per 10k and public worship attendants per 2k. On the right, histogram and boxplot of the variable of interest : the number of criminals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

To better visualize how the predictors are related with each other, we present them in a pairwise correlation plot along with their pearson correlation coefficients which were computed independently of the region code. From Fig.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.

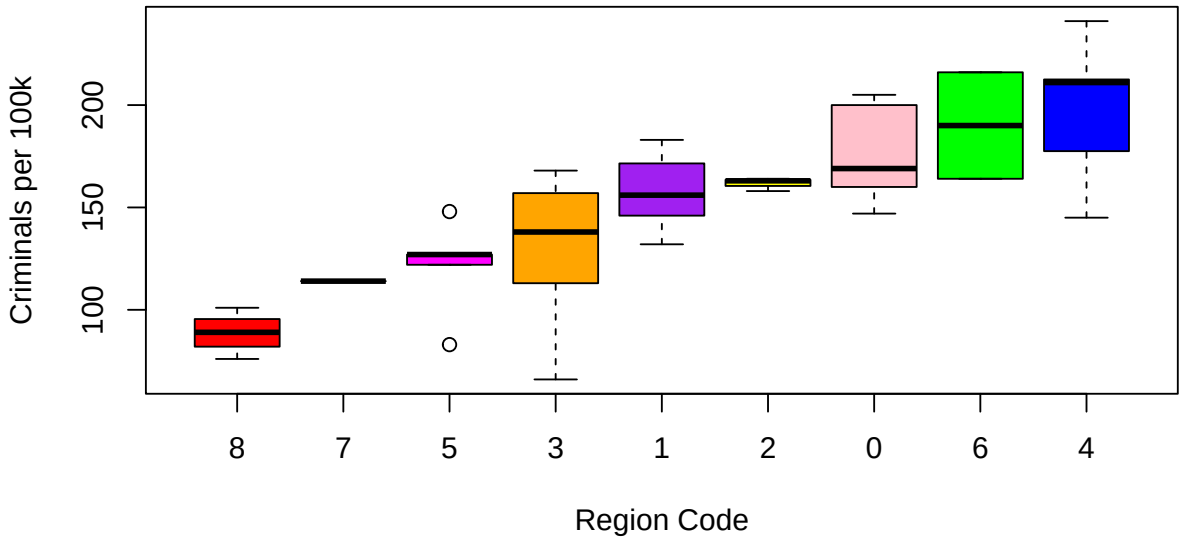

Figure 2 : Number of criminals per 100k by region code. Colors of the region code remain the same for the whole analysis.

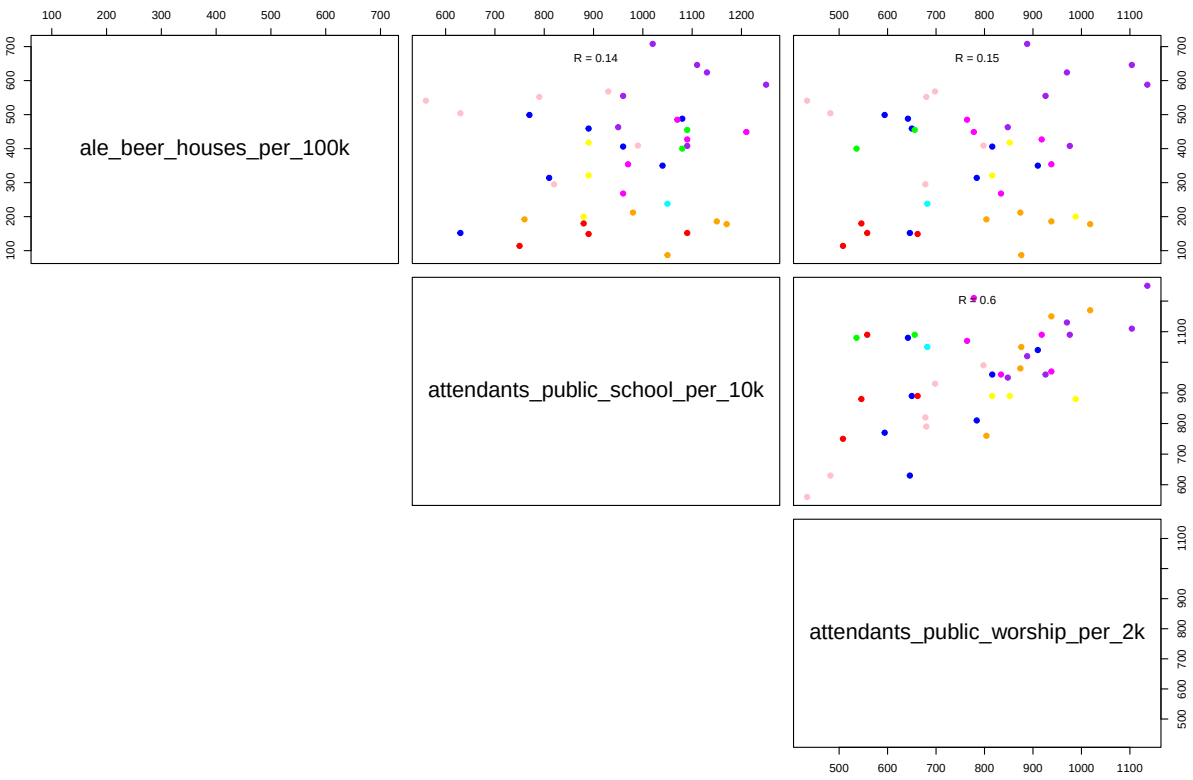

Figure 4: Pairwise correlation plot between the three predictors : Ale/Beer Houses per 100k, public school attendants per 10k and public worship attendants per 2k. Pearson correlation coefficients are computed independently of the region code. Colors correspond to region codes as in Figure 2.

Model Fitting and selection

In the following section, we will run multiple models and compare them via the adjusted R squared (hereafter R^2_{adj}) and the Akaike Information Criterion (hereafter AIC) metrics. As opposed to R^2 which is computed as $R^2 = \frac{\text{SSRegression}}{\text{SSTotal}} = 1 - \frac{\text{SSError}}{\text{SSTotal}}$, the adjusted R^2 penalizes complex models with more predictors through a penalty. We have $R^2_{adj} = 1 - \frac{\text{MSError}}{\text{SSTotal}/(n-1)} = 1 - \frac{n-1}{n-1-k}(1-R^2)$ where k is the number of predictors and n the sample size. The AIC also aims at penalizing complex models, and is computed as $\text{AIC} = -2\log L + 2p$ with L the maximum likelihood of the model and p the number of parameters of the model, combining a measure of deviance and a measure of complexity in order to obtain a good balance between the two. These two metrics will help us determine which model is the best to explain the number of criminals. We start with a model including all predictors and then successively change which variables are included in the model, basing our thought process on previous observations, and all results of the mentionned models can be found in Table 4.

Our base model has a R^2_{adj} of 0.26 and an AIC of 405.05, which we will use as a reference for the following models. In the exploratory data analysis, we saw that the number of public school attendants and the number of public worship attendants were positively correlated. Therefore, we first decided to remove the latter in the next model.
The second model, without public worship attendants, has a R^2_{adj} of 0.26 and an AIC of 404.041, which is better than the base model as the R^2_{adj} is higher and the AIC is lower. Next, we will see if the number of public worship attendants contributes more to the performance of the model compared to the number of public school attendants and leads to better results. The third model, without public school attendants, has a R^2_{adj} of 0.18 and an AIC of 408.52, which is worse in both R^2_{adj} and AIC than the second model. This indicates that the number of public school attendants contributes to the performance of the model and should be kept. To make sure that the number of public school attendants really contributes to the improvement of the performance, we decided to remove it in the next model, leaving only the number of ale and beer houses. The fourth model, only with ale and beer houses, has a R^2_{adj} of 0.19 and an AIC of 406.75, which is worse in R^2_{adj} and AIC than the second model. This indicates that the number of attendants at public schools contributes to the performance of the model and should be kept. Finally, we wanted to make sure that the number of ale and beer houses was also required for the model to perform well. To do that, we only consider the number of public school attendants in the last model, and find that the R^2_{adj} is 0.15 and the AIC is 414.22, which is worse than the second model, thus both the number of ale and beer houses and the number of public school attendants are needed to have the best performing model. To conclude, the best model is the second one, which includes both the number of ale and beer houses and the number of public school attendants. This model has a R^2_{adj} of 0.26 and an AIC of 404.05, which are the best values among all the models we considered, and its

equation is $\hat{Y} = 178.81 + 0.13 \cdot x_0 - 0.077 \cdot x_1$ where Y is the number of criminals per 100k, x_0 is the number of ale and beer houses and x_1 is the number of public school attendants.

Model	R^2_{adj}	AIC
Base model	0.26	405.05
Without public worship	0.26	404.05
Without public school	0.18	408.52
Only with ale and beer houses	0.19	406.75
Only with public school	0.15	414.22

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

Model Assessment

In the last section, we computed different models to determine which one is better. However, before accepting our model, it is essential to verify that the model's assumptions hold. As mentioned in an online article by Zach Bobbitt (Bobbitt 2020), the assumtions are that the residuals are normally distributed homoscedastic and independent, and that there is a linear relationship between the predictors and the response variable. We will not check the last assumption, as it is the base of the linear regression, however we will check the three others.

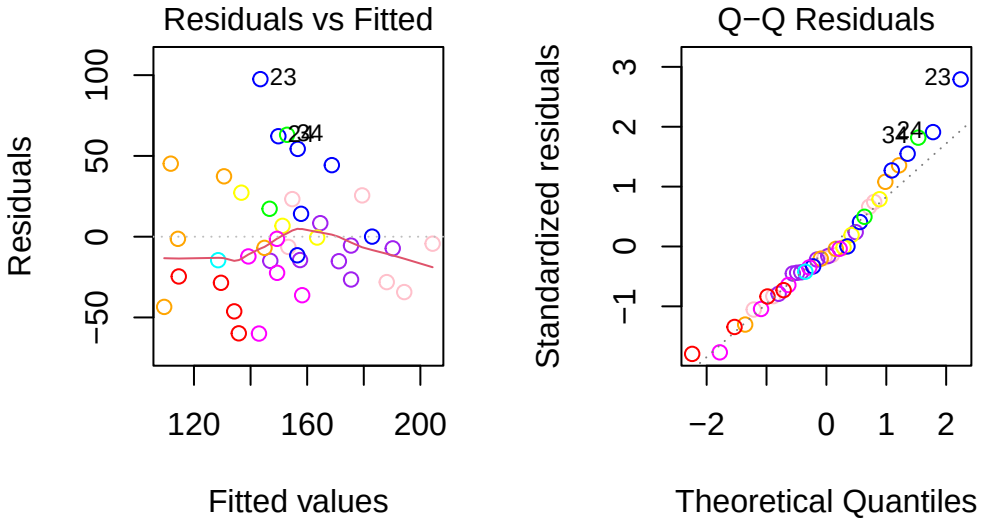

Figure 5: Diagnostic plots of the final model : Residuals vs Fitted plot (left), and QQ-normal plot of residuals (right). Colors correspond to region codes as in Figure 2.

By looking at the above 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.

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, and when 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, while the number of attendants to public school seemed to be the same. We also noticed that there was a strong positive correlation between the number of people attending public school and the number of people attending public worship.

Based on these observations, we ran multiple linear regression models to determine which variables were the best to explain the number of criminals and in the end we found that the best model, chosen by considering both the explained variance (R^2_{adj}) and the AIC, was $\hat{Y} = 178.81 + 0.13 \cdot x_0 - 0.077 \cdot x_1$, with Y the number of criminals, x_0 the number of ale and beer houses and x_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, most probably due to the fact that our initial hypotheses were based on data from two particular counties only, while the final model considers all of them. 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 therefore may not be able to fully explain the number of criminals in those regions.

Appendix

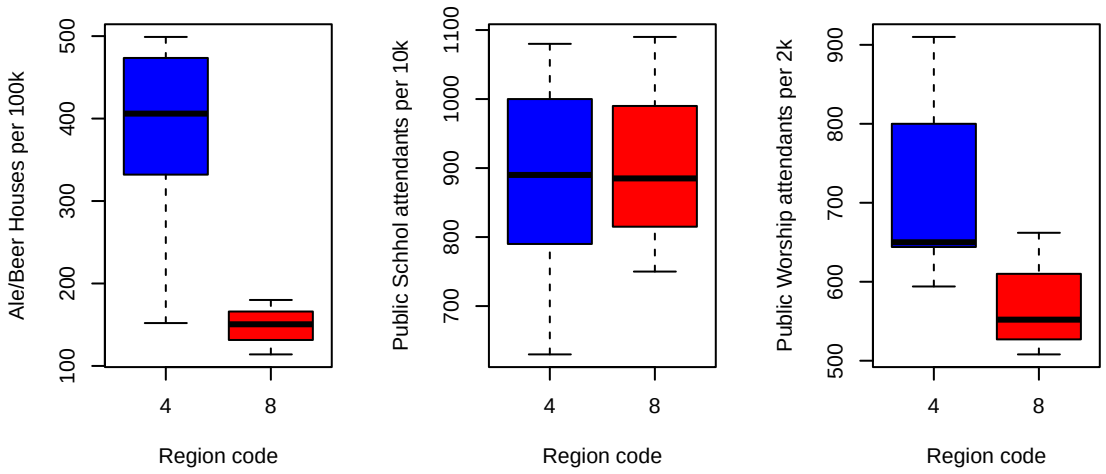

Figure 3 : Number of ale/beer Houses per 100k, public School attendants per 10k and public Worship attendants per 2k (the three predictors) for region Codes 4 and 8.

Bibliography

Bobbitt, Zach. 2020. "The Four Assumptions of Linear Regression — Statology.org." https://www.statology.org/linear-regression-assumptions/.