Analysis of Crime Predictors

Arvind Menon, Lennart Platon Kutzschebauch, Nisrine Bachar, Jayati Sood

11 July, 2024

Introduction

We as human beings categorize actions as moral and immoral, and label serious moral transgressions such as murder, theft and fraud as crimes. To prevent crime, we impose punishments as deterrents, though other methods to reduce crime may exist. Moreover, crime rates vary by region, prompting the question: what makes a crime more likely to happen in a region?
Clay (1857), concentrating on crime in England, analysed the following characteristics: Beer-(ale)-house numbers, worship attendance and public school attendance. The argument he provides for the latter two, is both allow for the public access to Christian moral values. It is to note, the "public schools" mentioned by Clay (1857) concentrated mostly on teaching read and writing and were not free but accepted anyone who could afford it. In the case of beer-houses, he argues "the temptation to animal pleasure" corrupts a person. To help his case he tries to show the positive correlation of beer-houses, the negative correlations of public school and worship attendance to crime.
In this paper we want to investigate these effects of the aforementioned characteristics and try to predict crime rate from them using the same data set. However, before starting it is important to mention, as noticed by Clay (1857), the recorded crime rate may be inaccurate because of systematic reasons since each county handles crime and punishment differently on the executive and juridical levels. This will impact the accuracy of all possible analysis and models.

Exploratory Data Analysis

We begin our analysis with a look at the first few lines of our dataset.

County	RegionName	RegionCode	Criminals_100k	BeerAle_100k	SchoolAttendance_10k	WorshipAttendance_2k
Middlesex	SouthEastern	1	200	541	560	43
Surrey	SouthEastern	1	160	504	630	48
Kent	SouthEastern	1	160	552	790	68
Sussex	SouthEastern	1	147	295	820	67
Hants	SouthEastern	1	178	409	990	79
Berks	SouthEastern	1	205	568	930	69

Data summary:

County	RegionName	RegionCode	Criminals_100k	BeerAle_100k	SchoolAttendance_10k	WorshipAttendance_2k
Length:40	Length:40	Min. :1.00	Min. : 66.0	Min. : 87.0	Min. : 560.0	Min. : 43.0
Class :character	Class :character	1st Qu.:1.00	1st Qu.:127.0	1st Qu.:209.0	1st Qu.: 880.0	1st Qu.: 65.0
Mode :character	Mode :character	Median :3.00	Median :157.5	Median :407.0	Median : 965.0	Median : 79.5
NA	NA	Mean :3.45	Mean :152.9	Mean :374.9	Mean : 957.8	Mean : 77.5
NA	NA	3rd Qu.:5.00	3rd Qu.:174.2	3rd Qu.:490.8	3rd Qu.:1082.5	3rd Qu.: 91.0
NA	NA	Max. :8.00	Max. :241.0	Max. :708.0	Max. :1250.0	Max. :113.0

The dataset contains information about 40 different counties.County and RegionName are categorical variables, and each of the 8 regions is assigned a RegionCode, which is a number from 1 to 8. Criminals_100k is the number of criminals per 100,000 inhabitants for any particular county. Similarly, BeerAle_100k, SchoolAttendance_10k and WorshipAttendance_2k are social indicators measured numerically as a proportion of the population. In order to better visualize the numerical data, we plot histograms of each numerical variable, with the exception of the categorical RegionCode.

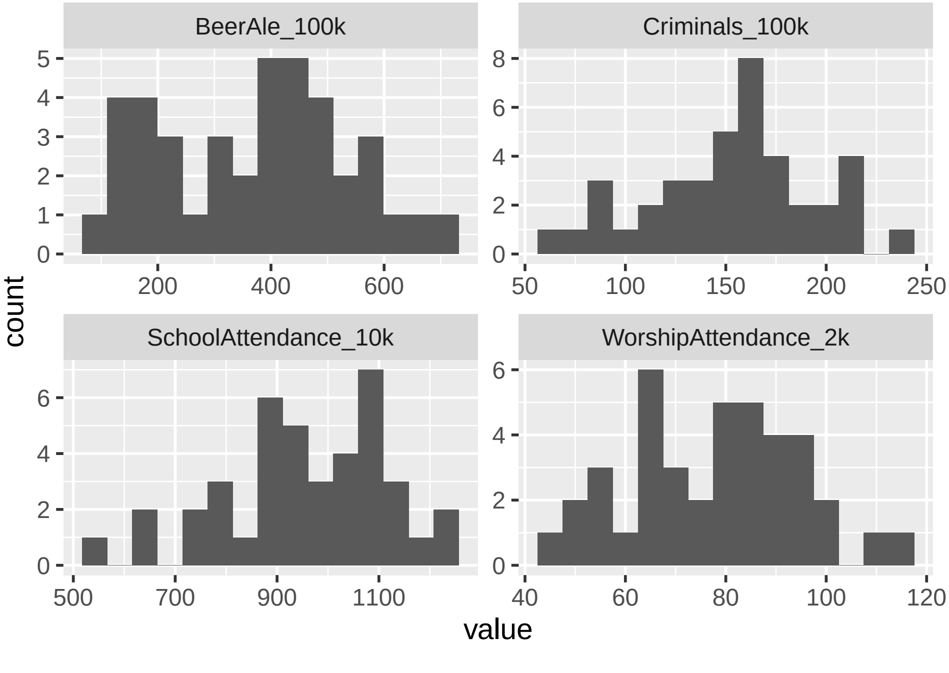

Figure 1: Histograms of numerical variables

The BeerAle_100k variable appears to be bimodal. The histogram suggests that there are two common levels of alcohol consumption within the entire population, one around the 200, and another around 400 per 100,000 population. According to the data summary, Crimanals_100k has a mean of 152.9 and a median of 157.5, suggesting the symmetry that is also reflected in the histogram. SchoolAttendance_10k is slightly left skewed, while WorshipAttendance_2k has a varied but loosely symmetric distribution.

The linear dependence between each pair of numerical variables is expressed in the following correlation matrix:

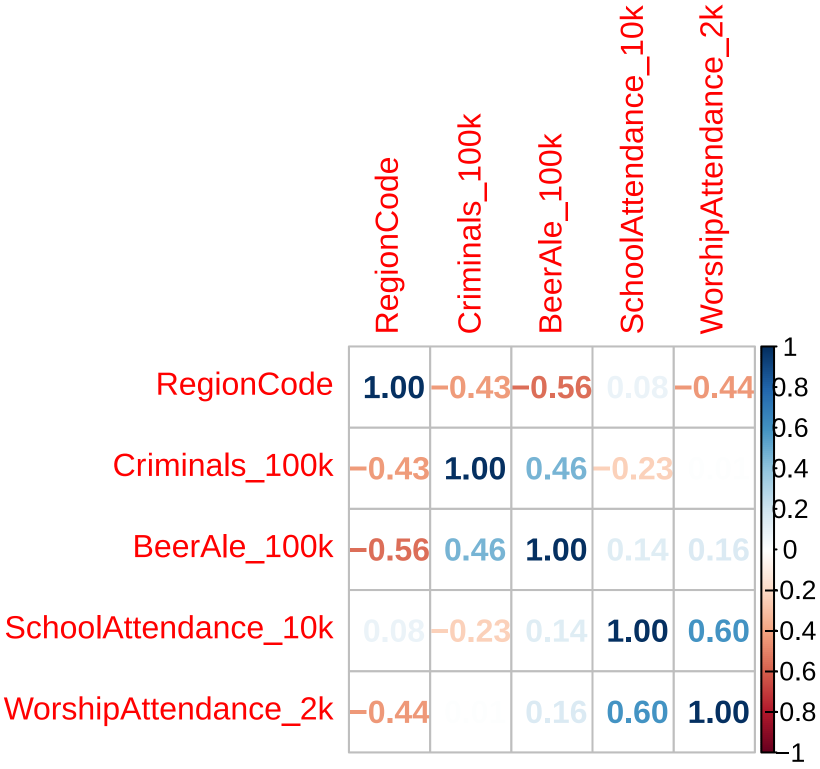

Figure 2: Correlation matrix

There appears to be negligible linear dependence between worship attendance and criminality. School attendance is slightly negatively correlated with the prevalence of crime. Criminal behavior is positively correlated with BeerAle_100k with a correlation coefficient of 0.46, suggesting that counties with a more dominant culture of frequenting bars and pubs are also where more crime happens. In order to better visualize these dependencies, we regress Criminals_100k on each of these variables.

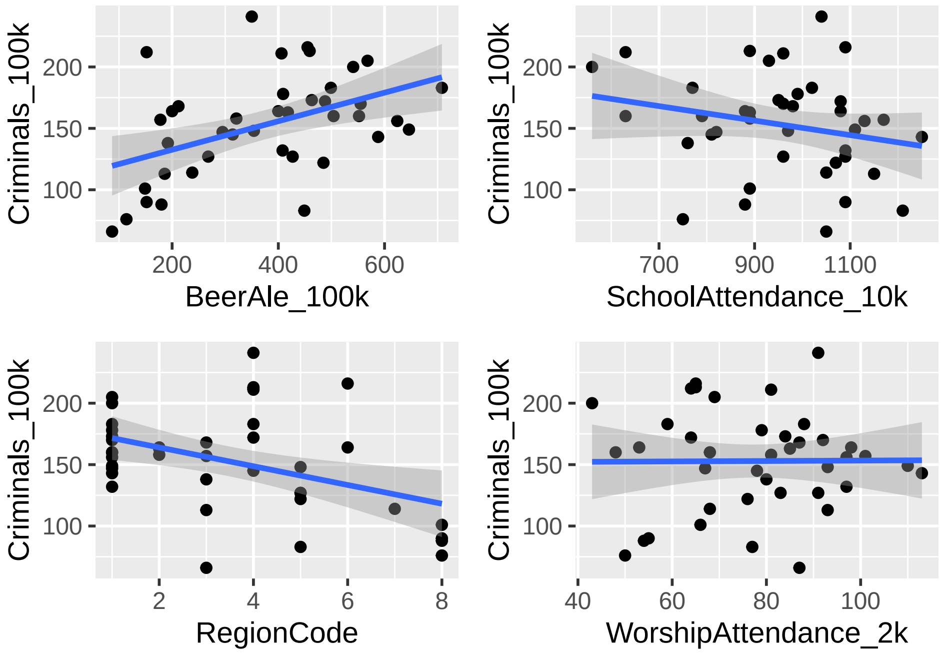

Figure 3: Regression plots of Criminals per 100k on the other variables

There appears to be a very clear linear dependence of criminal behaviour on bar attendance, with most data points falling within the 95% confidence interval of the regression line. There is high variation of criminality across school attendance and worship attendance.

Model Fitting

We use least squares to minimize the sum of squared residuals in a polynomial regression. In order to predict Criminals_100k, we use a polynomial regression model which uses BeerAle_100k, SchoolAttendance_10k, and WorshipAttendance_2k as the features. We include the variable WorshipAttendance_2k in our analysis even though it has no correlation with the variable Criminals_100k in order to capture any non-linear association.

A polynomial regression model of degree n with three features (x_1, x_2, x_3) can be represented as follows:

$$y = \beta_0 + \beta_{11}x_1 + \beta_{12}x_1^2 + ... + \beta_{1n}x_1^n + \beta_{21}x_2 + \beta_{22}x_2^2 + ... + \beta_{2n}x_2^n + \beta_{31}x_3 + \beta_{32}x_3^2 + ... + \beta_{3n}x_3^n + \epsilon$$

Where y is the predicted Criminals_100k population, and x_1, x_2, and x_3 represent BeerAle_100k, SchoolAttendance_10k and WorshipAttendance_2k respectively. β_0, β_{ij} are the coefficients of the model where i is the feature index and j is the degree of the polynomial ϵ is the error term.

We choose the model based on the exploratory data analysis and the correlation between the features and the target variable. Polynomial terms are included to capture any non-linear relationships between the features and the target variable. We use leave-one-out cross validation to find the optimum polynomial degree among polynomials of degrees up to 6. The minimum cross validation errors corresponds to polynomials with degrees 1, 3 and 4.

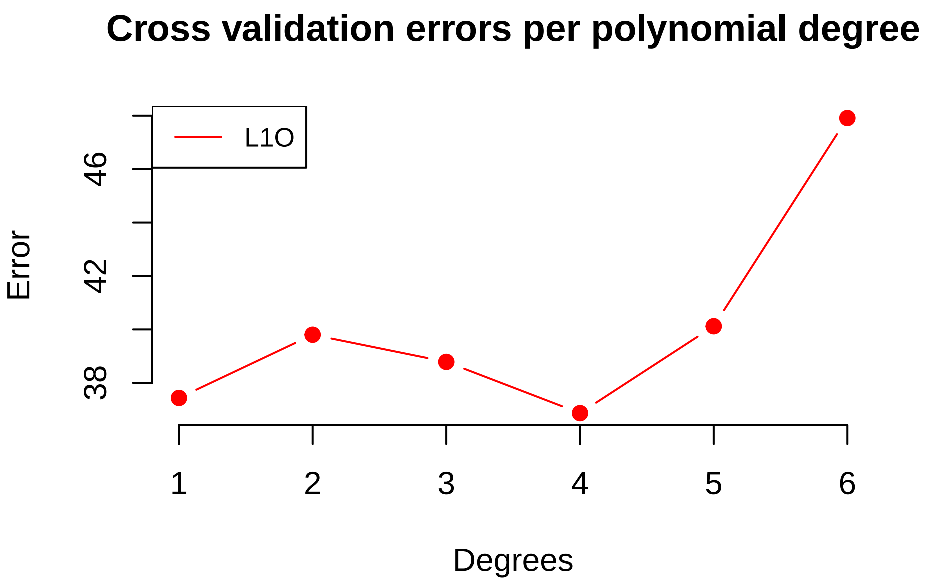

Figure 4: Cross validation errors

We further compare polynomial regression models of degrees 3 and 4 with the linear model using the Akaike Information Criterion (AIC) and the Bayesian Information Criterion. These are both measures to compare different models in order to determine which fits the data best, while penalizing the complexity of the model. They are calculated based on the number of parameters in the model and the likelihood of the model. The respective mathematical formulations are as follows:

$$\text{AIC} = 2k - 2\ln(L)$$

$$\text{BIC} = k\ln(n) - 2\ln(L)$$

where in both cases, k is the number of parameters and L is the maximum likelihood of the model. n is the number of observations. As we can see, BIC penalizes the number of parameters more strictly, and in proportion to the number of observations. It thus yields sparser models with fewer predictors.

Table 3: Information Criteria

	DF	AIC	BIC
Linear model	5	405.0249	413.4693
Degree 3 polynomial	11	402.5364	421.1140
Degree 4 Polynomial	14	401.7266	425.3710

According to AIC, the higher degree polynomial models fit better the data than the simpler linear model, but by a small margin. The BIC of the linear model is the lowest, indicating that a slightly stricter penalty on the number of parameters (ln(40)~3.69) yields the linear model as the better model. Given the test results and the regression estimates, we favour parsimony of predictors, and choose the linear model for further analysis.

Table 4: Regression Model Summary

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	172.927	35.807	4.829	0.000
BeerAle_100k	0.123	0.035	3.517	0.001
SchoolAttendance_10k	-0.101	0.044	-2.305	0.027
WorshipAttendance_2k	0.398	0.414	0.962	0.342

We get the following model:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Where $\hat{y}$ is the predicted Criminals_100k population, and x_1, x_2, and x_3 represent BeerAle_100k, SchoolAttendance_10k and WorshipAttendance_2k respectively. The regression coefficients β_0 and β_{ij} are estimated in Table 4. ¨

Model assessment

In this section, we will verify the conditions for our regressions model to hold, namely that our residuals have zero mean, are uncorrelated, are homoscedastic and are normally distributed.

Zero mean error terms

We compute our model residuals mean and get : $-2.0428104 \times 10^{-15}$, which is very close to 0 as needed.\

Homoscedastic error terms

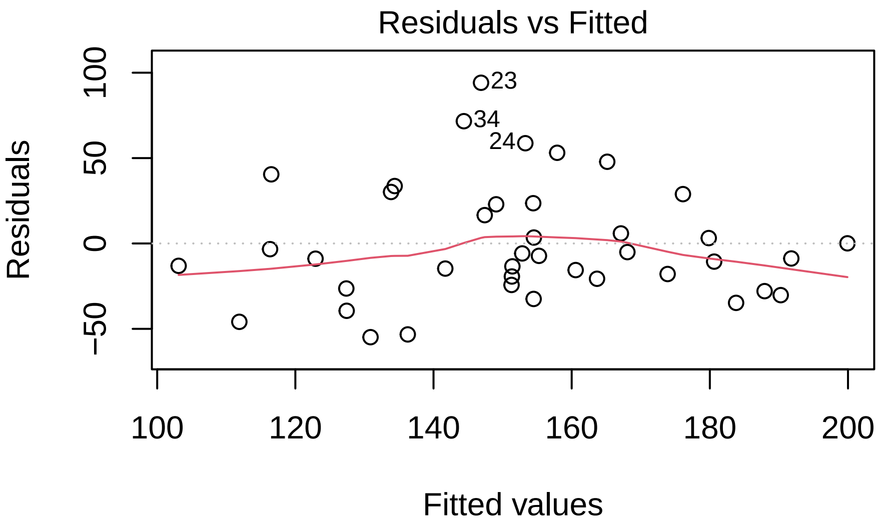

Figure 5: Residuals vs fitted values plot

We plot residuals against fitted values and find that our residuals seem to be uniformly scattered about the mean 0. There appears some clustering in the center, however, this is to be expected from a uniform distribution and the amount of data points we posses. This implies that the data is almost homoscedastic.

Correlation between error terms

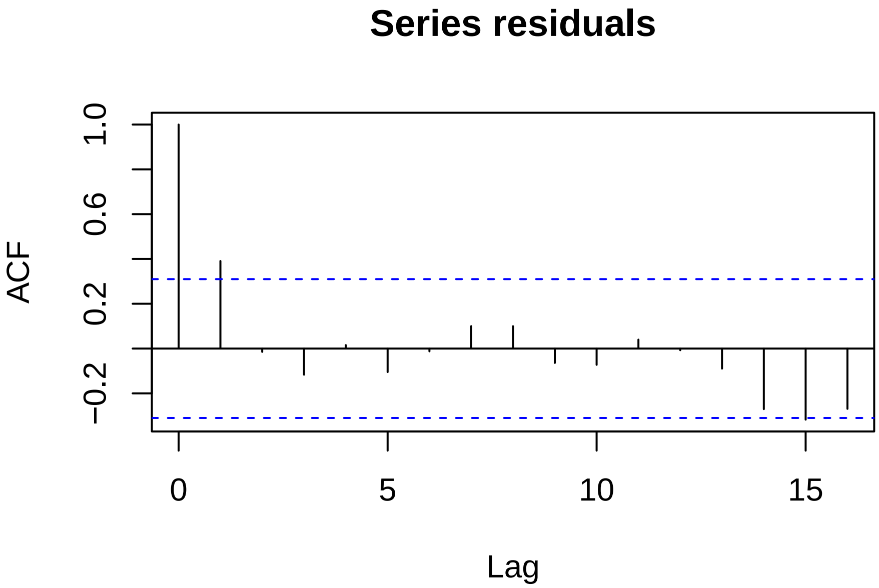

Figure 6: Residuals Autocorrelation function

The Autocorrelation Function (ACF) plot displays the correlation between the residuals at different lags, with the vertical bars representing the autocorrelation coefficients and the horizontal blue dashed lines representing 95% confidence intervals for the coefficients. At lag 0, the autocorrelation is always 1 as expected. For other lags, since most of the bars are within the 95% confidence interval band, we conclude that the residuals have little to no significant autocorrelation for most lags. Since the bar at lag 1 extends slightly beyond the confidence interval, there may be minor autocorrelation at lag 1, which we confirm by calculating its Durbin-Watson statistic.

Both the autocorrelation function plot and the Durbin-Watson test (with autocorrelation and p-value) suggest that we have a weak positive autocorrelation at lag 1. It may be due to an omitted variable bias, as it is specified in our reference paper that criminality and especially sentences depend of the region since if it is crowded the law tends to be more strictly applied.

Distribution of error terms

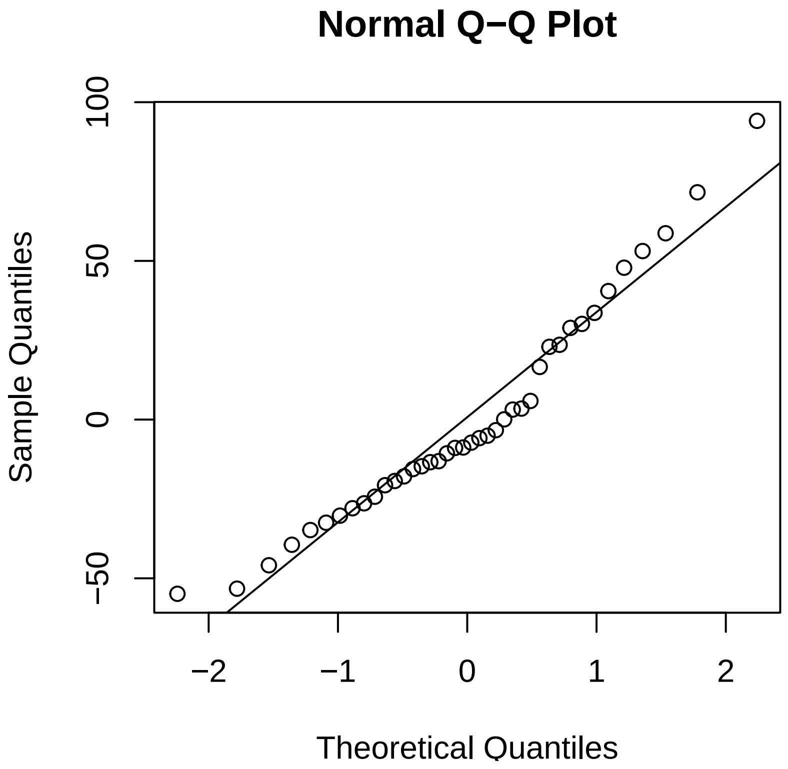

Figure 7: Q-Q plot

We notice the points in the QQ-plot mostly follows the diagonal with some discrepancies at the lower end. Given that our data set is small, we can assume from the Normal QQ plot that our residuals are approximately normally distributed.

Overall, we tend to see some minor indication that not all conditions needed for the models are perfectly met. However, given the amount of data points and that the deviations are not too noticeable we believe that all conditions are approximately met and our model fitting is robust enough but some biases might be present.

Conclusion

Our analysis begins with exploratory data analysis, and it made some key observations. "Criminals per 100k" had a positive correlation(=0.46) with "Ale/Beer houses per 100k", a negative correlation(=0.23) with "Attendants at school per 10k", and a very insignificant correlation with "Attendants at public worship per 2k". To investigate non-linear relations, we choose a polynomial regression model and compared its performance with that of a linear regression model. We used leave-one-out cross validation to select polynomials of degrees 3 and 4 for our polynomial regression model. Models with degrees 3,4 both have a lower AIC than the linear regression model indicating a better fit, of which we finally chose a degree 3 model, based on the results from the likelihood ratio test.

The fitted model has positive coefficients for “Ale/Beer houses per 100k” features, negative coefficients for “Attendants at school per 10k” features, and opposite sign coefficients for “Attendants at public worship per 2k”, which indicates that the features cancel each other out and end up being less significant. These results align with our initial analysis and partially agree with the claims made in the study, with the difference in the significance of “Attendants at public worship”, which is found to be insignificant. On analysing the residual errors, we find that they are normally distributed (using the Q-Q plots) and almost homoscedastic (from the residual vs fitted values plot). The error terms are uncorrelated with a weak positive autocorrelation at lag 1. Thus the model adheres to the assumptions made during model selection.

In summary, the model agrees with the study that the crime rate in a county is positively associated with the density of ale/beer houses in that county, and negatively associated with the density of attendants at schools in that county, and disagrees with the fact that the density of attendants at places of public worship in that county has a significant influence on the crime rate in that county.

References

Clay, John. 1857. “On the Relation Between Crime, Popular Instruction, Attendance on Religious Worship, and Beer-House.” Journal of the Statistical Society of London 20 (1): 22–32. http://www.jstor.org/stable/2338159.