Criminals Population Prediction

Laurine Song, Marine Ract, Mathilde Roulleau, Julie Le Flem

Introduction

One could imagine that attending school or practicing a religion might reduce criminality, and that alcohol consumption might increase it.

Using proxy for those behavior —respectively the proportion of population attending school and worship, and the proportion of bars per habitant— we attempt to model criminality as a function of these proxy using linear regression.

The dataset we are using is extracted from a statistical journal (Clay 1857), which includes data from 40 different English counties taken over 5 years ending in 1853. The following features were given for each county:

- County
- Region name
- Region code
- Criminals (per 100k)
- Beerhouses (per 100k)
- School attendance (per 10k)
- Worship attendance (per 2000)

Research question

Is there a relation between crime, popular instruction, attendance on religious worship, and beer-house? What could be changed to reduce criminality in these counties ? What influences the criminality in these counties ?

Exploratory Data Analysis

For this statistical analysis, we select the 4 last features, load them and normalize them to get all the numbers per 100. We then check that there is no missing value (Table 1).

To start the exploration of the data, we perform a numerical and graphical univariate analysis for the 4 continuous variables (Table 2, Figure 1). The distribution of criminals per 100 inhabitants appears right-skewed (Figure 1.1), meaning most areas have a low to moderate number of criminals per 100 people. The distribution of beer houses is more spread out (Figure 1.2), suggesting variability in the number of beerhouses per 100 people across locations. There seem to be two peaks (bimodal distribution), indicating that some areas have relatively few beerhouses while others have many. The distribution of school attendance shows a somewhat normal distribution (Figure 1.3), with most values concentrated around 9 to 11 schools per 100. It suggests that the number of schools per 100 people is relatively stable across different areas. The distribution of worship attendance

presents multiple peaks spread out between 20 and 60 (Figure 1.4), suggesting a high variability of religious density across areas.

We then perform a numerical and graphical bivariate analysis for the 4 continuous variables (Table 3, Figure 2). We plot the pairwise correlations of all variables and compute the numerical strength of association between all pairs of variables with Pearson's correlation coefficients. Criminals are positively correlated with beer houses (0.46) and show almost no correlation with school and worship. The two predictors school and worship are highly positively correlated (0.60).

Model Fitting

Given that we are predicting criminals per 100 inhabitants as a function of other variables, we consider a multiple linear regression:

$$Y = \beta_0 + \beta_1 * X_1 + ... + \beta_n * X_n$$

with Y: Dependent variable, X_n: Predictors, and β_n: Estimated coefficients.

We want to develop the best predictive equation for criminality based on the 3 predictors: instruction, attendance on religious worship, and the beer-houses. To do so, we select the best model using the Forward selection approach, a stepwise regression technique that starts with an empty model and iteratively adds the most significant predictor until no further improvement is possible. At each step, the Akaike Information Criterion (AIC) is computed until the lowest value is reached. This approach fits the models using Ordinary Least Squares (OLS). The process starts with an empty model (Criminals 1).
The first variable added is Beerhouses, which significantly reduces the AIC from -253.74 to -261.38. The second variable added is School, which further reduces the AIC to -264.10. Adding Worship does not improve the model significantly (AIC only decreases to -263.09), so it is not included in the final model (Table 4).

We obtain a model (Table 5) with significant coefficients, all associated with a p-value $<$ 0.05. The adjusted R-squared, equal to 0.26, is low meaning that only 26% of the variance is explained by the independent variables.

Model Assessment

In our selected model, we ideally want the errors to have mean 0, to be homoscedastic (same variance), uncorrelated and normally distributed. To carry out these assessments, we plot the residuals vs. fitted, the QQ normal plot of residuals, the scale-location plot and the residuals vs. Leverage.

- The red smoothed line suggests slight curvature, which may indicate non-linearity. The spread appears somewhat uneven, hinting at possible heteroscedasticity (Figure 3.1)
- There is deviation at the upper end (right tail), meaning some residuals are not normally distributed, which could indicate outliers or skewness (Figure 3.2).

- Some variation in spread is present, which may indicate heteroscedasticity (Figure 3.3).
- Some points, like "240" and "230," are more distant, suggesting potential influential observations (Figure 3.4).

These 4 plots suggest that our model presents issues of non-linearity, heteroscedasticity and non-normality residuals as well as some influential points.

Final estimated model

The linear model that best predicts the number of criminals in the English counties as a function of beer houses and school attendance is defined as:

$\widehat{Criminals} = 0.18 + 0.13 * Beerhouses - 0.01 * School$

Plots

	Missing Values
criminals_per_100	0
beerhouses_per_100	0
school_per_100	0
worship_per_100	0

Table 1 : Missing values in the data.

criminals_per_100	beerhouses_per_100	school_per_100	worship_per_100
Min. :0.0660	Min. :0.0870	Min. : 5.600	Min. :21.70
1st Qu.:0.1270	1st Qu.:0.2090	1st Qu.: 8.800	1st Qu.:32.73
Median :0.1575	Median :0.4070	Median : 9.650	Median :40.05
Mean :0.1529	Mean :0.3749	Mean : 9.578	Mean :39.01
3rd Qu.:0.1742	3rd Qu.:0.4908	3rd Qu.:10.825	3rd Qu.:45.60
Max. :0.2410	Max. :0.7080	Max. :12.500	Max. :56.80

Table 2: Statistical summary of the data.

Figure 1: Distribution of the Variables

	criminals_per_100	beerhouses_per_100	school_per_100	worship_per_100
criminals_per_100	1.0000000	0.4628628	-0.2300452	0.0035630
beerhouses_per_100	0.4628628	1.0000000	0.1354865	0.1531569
school_per_100	-0.2300452	0.1354865	1.0000000	0.5968562
worship_per_100	0.0035630	0.1531569	0.5968562	1.0000000

Table 3: Pearson's correlation coefficients.

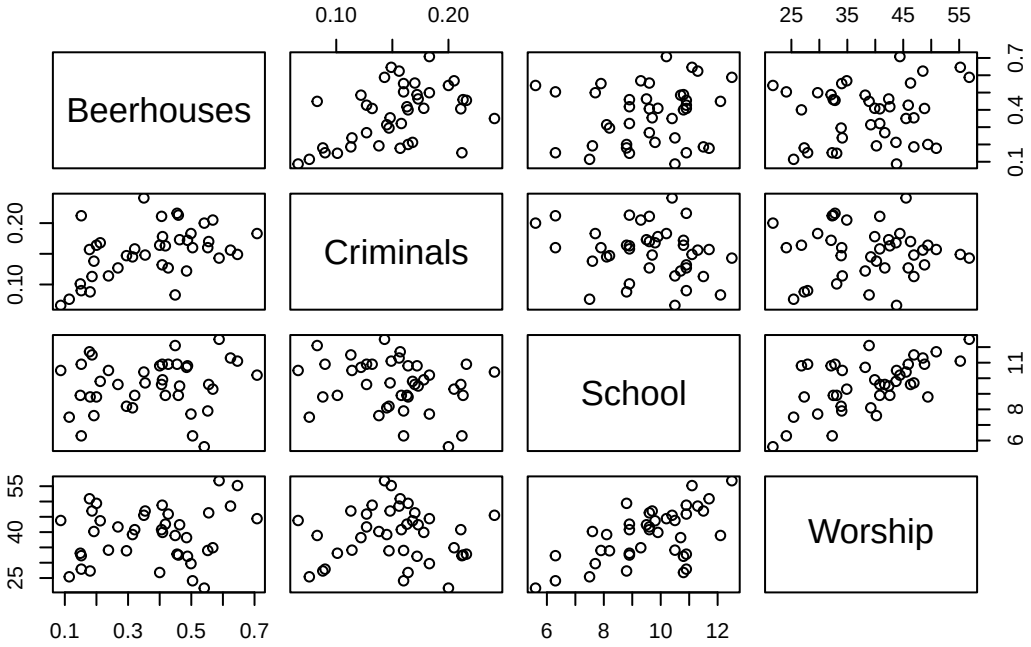

Figure 2: Pairwise Correlation of the variables.

```
Start:  AIC=-253.74
Criminals ~ 1

             Df Sum of Sq      RSS     AIC
+ Beerhouses  1 0.0143323 0.052565 -261.38
+ School      1 0.0035403 0.063357 -253.91
<none>                    0.066898 -253.74
+ Worship     1 0.0000008 0.066897 -251.74

Step:  AIC=-261.38
Criminals ~ Beerhouses

          Df Sum of Sq      RSS     AIC
+ School   1 0.0058408 0.046725 -264.10
<none>                 0.052565 -261.38
+ Worship  1 0.0003105 0.052255 -259.62

Step:  AIC=-264.09
Criminals ~ Beerhouses + School

          Df Sum of Sq      RSS     AIC
<none>                 0.046725 -264.10
+ Worship  1 0.0011439 0.045581 -263.09
```

Table 4: Forward model selection.

```
Call:
lm(formula = Criminals ~ Beerhouses + School, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.05997 -0.02293 -0.00592  0.01877  0.09753

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.178813   0.035246   5.073 1.12e-05 ***
Beerhouses   0.126351   0.034815   3.629 0.000854 ***
School      -0.007651   0.003557  -2.151 0.038108 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03554 on 37 degrees of freedom
Multiple R-squared:  0.3016,    Adjusted R-squared:  0.2638
F-statistic: 7.987 on 2 and 37 DF,  p-value: 0.001308
```

Table 5: Result of Forward Model Selection.

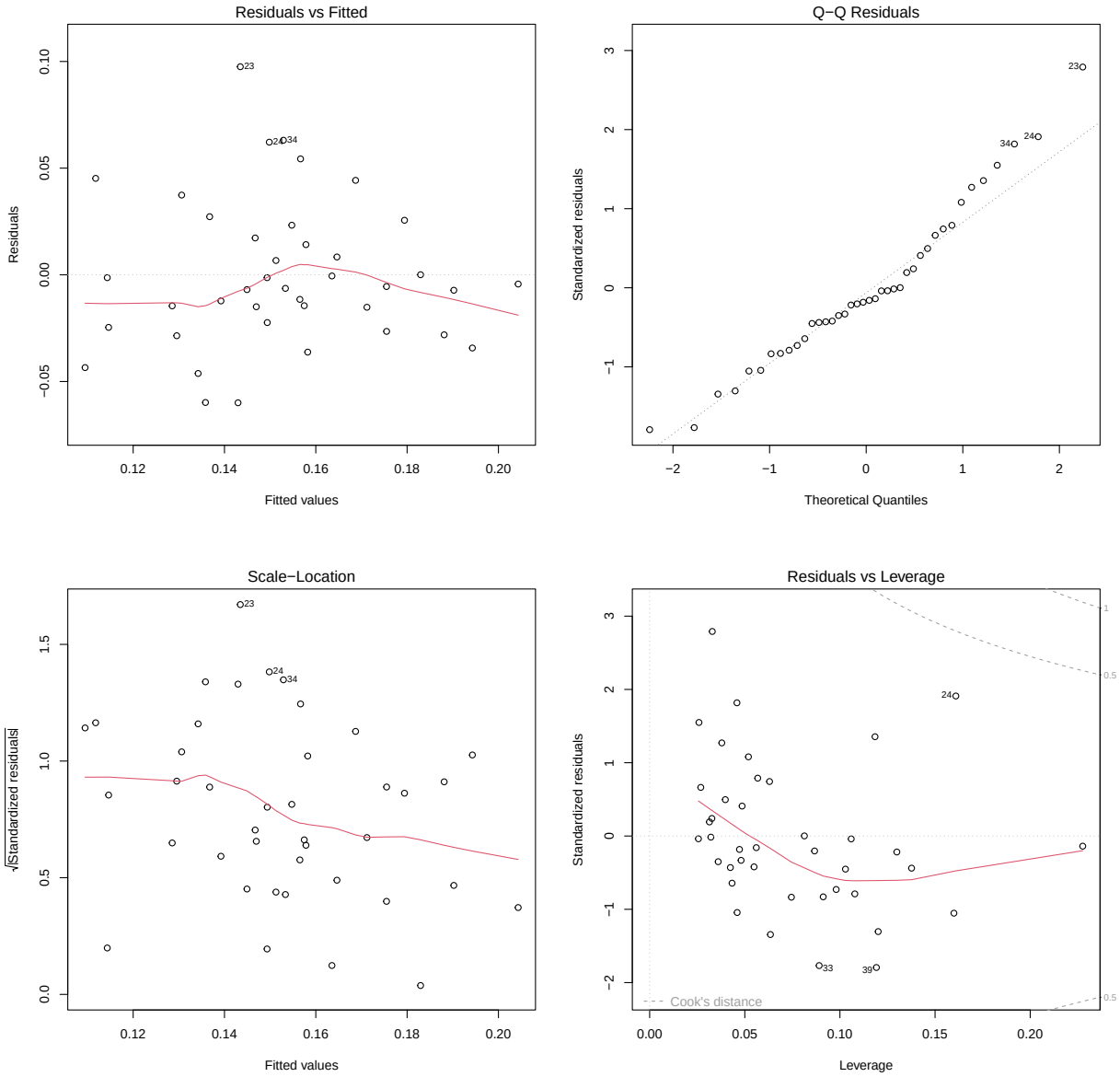

Figure 3: Diagnostic Plots.

Conclusions

To conclude, we managed to find a model that was able to predict the number of criminals in a county as a function of the number of beer houses and school attendance. On average, in British counties circa 1850, there were 180 criminals per 100k population. More beer houses make the number of criminals increase while more school attendance makes it decrease. To illustrate, approximately 8 additional beerhouses per 100k pop would lead to 1 additional criminal. Controversy, 100 additional school attendants would reduce the number of criminals by 1. We also found that the worship attendance was not a variable significant enough to predict the number of Criminals. The adjusted R-squared value for our model is acceptable, meaning a part of the variability of the observed data is explained by the chosen variables. However, they do not explain all the variability and it could be interesting to add more variables (e.g. profession, age, salary). It could also increase the accuracy and significance of our model by adding more data. It would allow us to exclude

some outliers we decided not to exclude as we did not have a sample large enough. It would also be better to fit the model assumptions.

References

Clay, John. "On the Relation Between Crime, Popular Instruction, Attendance on Religious Worship, and Beer-House." Journal of the Statistical Society of London 20, no. 1 (March 1857): 22-32. https://www.jstor.org/stable/2338159.