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

Predicting counties' Criminals Population based on Beerhouses, School and Worship attendance

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 [1], 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 relationship between the number of criminals of England's counties and popular instruction, attendance on religious worship, and number of beerhouse? To what extent do these features predict the number of criminals in 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. Before starting the exploration of the dataset, we first verified that there were no missing values in the dataframe. Then, we perform a numerical (Table 1) and graphical (Figure 1) univariate analysis for the 4 continuous variables.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
criminals_per_100	0.07	0.13	0.16	0.15	0.17	0.24
beerhouses_per_100	0.09	0.21	0.41	0.37	0.49	0.71
school_per_100	5.60	8.80	9.65	9.58	10.83	12.50
worship_per_100	21.70	32.72	40.05	39.01	45.60	56.80

Table 1: Statistical summary of the data.

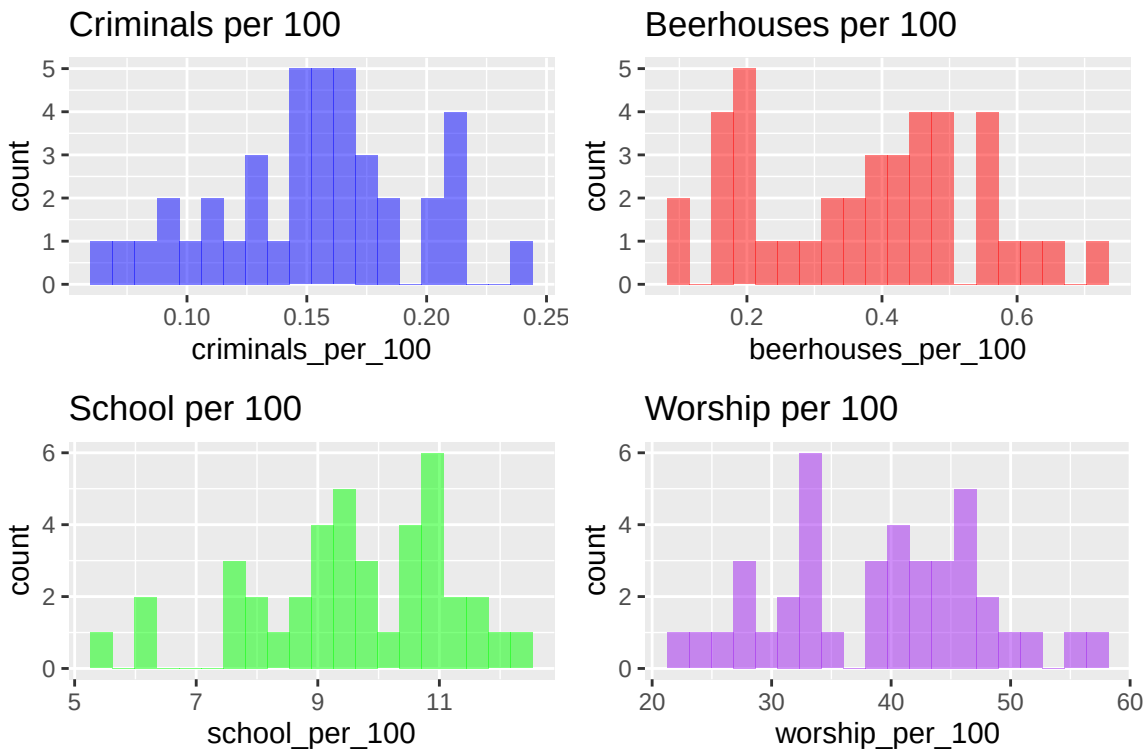

Figure 1: Distribution of the Variables

In Figure 1, the distribution of criminals per 100 inhabitants appears right-skewed, meaning most areas have a low to moderate number of criminals per 100 people. The distribution of beer houses is more spread out, 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, 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. At last, the distribution of worship attendance presents multiple peaks spread out between 20 and 60, suggesting a high variability of religious density across areas.

We then perform a numerical and graphical bivariate analysis for the 4 continuous variables (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. These coefficients quantify the strength and direction of the linear relationships between two continuous variables. For two variables X and Y, the Pearson correlation coefficient r is:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Where X_i, Y_i are individual observations, and $\bar{X}$, $\bar{Y}$ are the sample means of X and Y.

Therefore, criminals are positively correlated with beer houses (0.46) and show almost no or weak correlation with school and worship. Also, the two predictors school and worship are highly positively correlated (0.60).

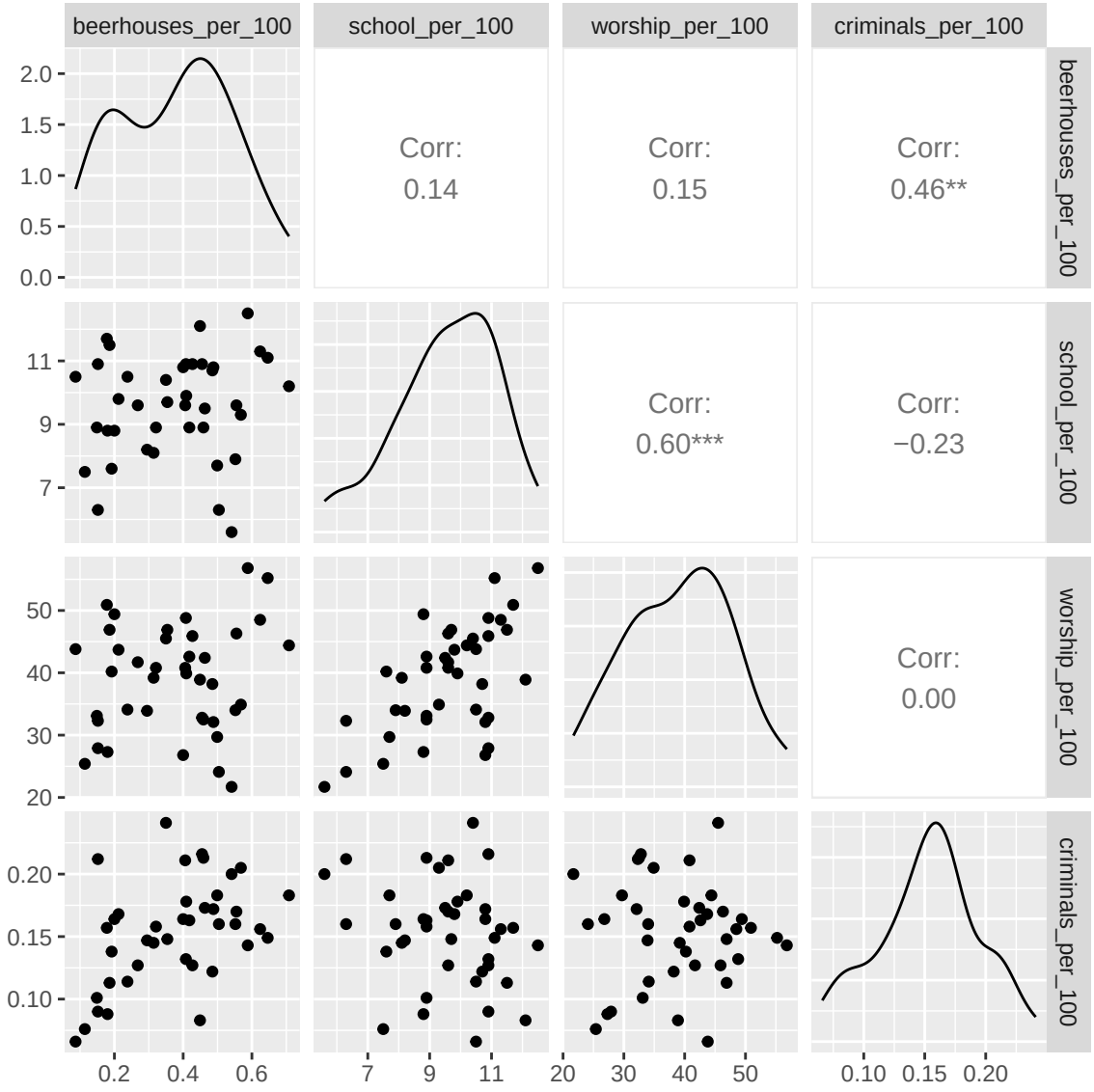

Figure 2: Pairplots of the variables with their Pearson's correlation coefficients.

Model Fitting

We aim to predict the number of criminals per 100 inhabitants as a function of other variables using a multiple linear regression:

$$C = \beta_0 + \beta_1 * X_1 + ... + \beta_n * X_n + \epsilon$$

with C the number of Criminals per 100 inhabitants, X_i: the predictors selected among instruction, attendance on religious worship, β_i the estimated coefficients for each corresponding predictors, and ϵ the error term.

To develop the best predictive equation for criminality based on the three predictors : instruction, attendance on religious worship, and beer-houses, we use the Forward Selection [2] approach. This is a stepwise regression technique that begins with an empty model and iteratively adds the most statistically significant predictor at each step, continuing until no further improvement in model performance is observed. The model selection at each stage is guided by the Akaike Information Criterion (AIC) [3], which balances model fit and complexity. The AIC is a measure of the relative quality of statistical models for a given dataset; it is defined as:

$$AIC = 2k - 2ln(L)$$

where k is the number of estimated parameters in the model, and L is the maximum value of the likelihood function for the model. A lower AIC indicates a better model by penalizing unnecessary complexity while rewarding goodness of fit. The process continues until the AIC cannot be further minimized. All models are fitted using Ordinary Least Squares (OLS) estimation.

The process starts with an empty model containing no predictors.

- 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.
- The third candidate variable, Worship, leads to an increase in AIC (–263.09). Worship is therefore not included in the final model.

This results in a final model that includes only Beerhouses and School as predictors of criminality, in which all coefficients are statistically significant, with associated p-values < 0.05 (Table 2). The residual standard error is 0.04, indicating that the typical deviation of the observed values from the predicted values is approximately 0.036 units, in the scale of the number of criminals. The model explains approximately 26% of the variance in the number of criminals, as indicated by an adjusted R-squared of 0.26. The model's overall significance is supported by an F-statistic of 7.99, with a corresponding p-value of 0.001. The multiple R-squared is 0.30, indicating 30% of variance in the number of criminal is explained by the model before adjusting for the number of predictors.

Variable	Estimate	Std. Error	t value	Pr(>|t|)	
Intercept	0.18	3.52e-2	5.07	1.12e-05	***
Beerhouses	0.13	3.48e-2	3.63	8.54e-04	***
School	-7.65e-3	3.56e-3	-2.15	3.81e-02	.

Table 2: Coefficient estimates after the forward model selection. The model is 'Criminals ~ Beerhouses + School'. Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Model Assessment

To evaluate the model, we plot the residuals vs. fitted, the QQ normal plot of residuals, the scale-location plot, and the residuals vs. Leverage.

- **Linearity:** We ideally want the relationship between predictors and the response to be linear. In Figure 3.1 (residuals vs. fitted), the red smoothed line shows a very slight curve, we consider the assumption to be approximately validated.
- **Mean of errors = 0:** We ideally want the errors to have a mean of 0. In Figure 3.1, this assumption appears to hold, as the residuals are scattered fairly evenly around the horizontal line at 0.
- **Homoscedasticity:** We ideally want the errors to have constant variance. In Figure 3.1, the spread of residuals appears somewhat even, clustered around 0, indicating homoscedasticity. This is echoed in Figure 3.3 (scale-location plot), in which residual are randomly scattered around the red line. Only a few points appear isolated from the red line in Figure 3.1 suggesting a slight heteroscedasticity.
- **Independence:** The errors of the model should be uncorrelated. Figure 3.1 shows no strong pattern or trend in residuals, which supports the assumption of independence.
- **Normality:** The residuals should be normally distributed. Figure 3.2 (QQ normal plot) only a few points show deviations from the reference line, particularly in the upper tail, suggesting a slight non-normality. Figure 3.4 also shows that certain points (e.g., “240” and “230”) may be influential and potentially contribute to non-normality.Overall the residuals are approximately normally distributed.

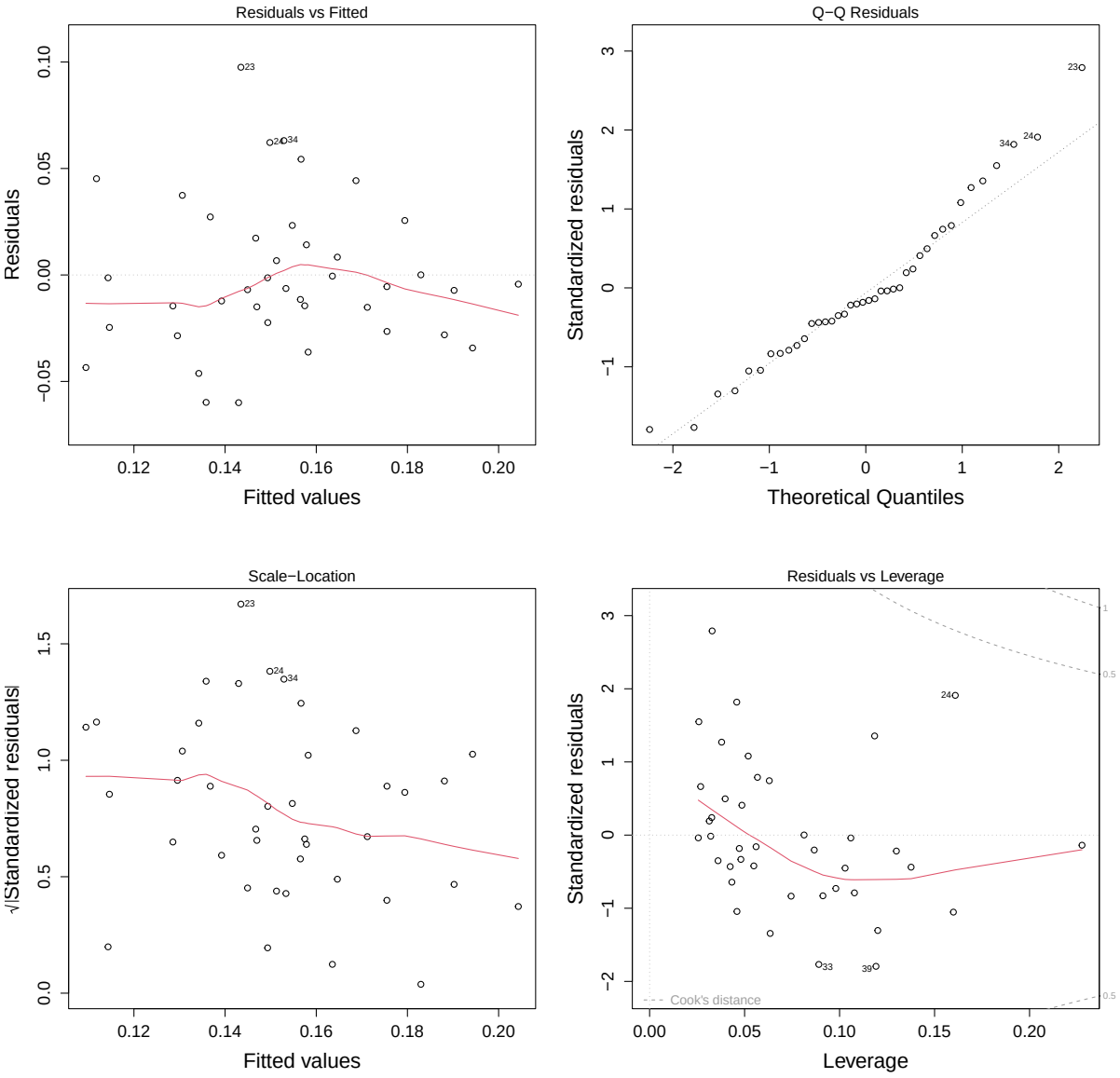

Figure 3: **Diagnostic Plots. 1) Residuals vs. Fitted**: Shows residuals against predicted values. Random scatter around the horizontal line at 0 indicates good fit. The red line shows the local average. **2) Normal Q-Q**: Compares standardized residuals to a normal distribution. Points should lie along the dashed diagonal if residuals are normally distributed. **3) Scale-Location**: Plots the square root of absolute standardized residuals vs. fitted values. A horizontal red line and random scatter suggest constant variance (homoscedasticity). **4) Residuals vs. Leverage**: Identifies influential points by plotting standardized residuals against leverage. Dashed Cook's distance lines highlight observations with high influence.

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:

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

For each additional beer house, the model predicts an increase of 0.13 criminals, assuming school attendance remains constant. This suggests a positive association between beer houses and crime. For each additional unit in school attendance, the model predicts a decrease of 0.01 criminals.This implies a negative association between school attendance and crime.

Conclusion

To conclude, we developed a linear model predicting 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 100,000 population. The number of beer houses is positively associated with criminality, while school attendance is negatively associated. Specifically, an increase of approximately 8 beer houses per 100,000 people is associated with one additional criminal, whereas an increase of 100 school attendants per 100,000 corresponds to a decrease of one criminal. Worship attendance did not significantly contribute to predicting criminality and was excluded from the final model.

The adjusted R-squared value of 0.26 indicates that the model explains about 26% of the variation in criminality, leaving a substantial portion unexplained. Including additional variables such as profession, age, or income might improve the model. Moreover, a larger dataset would enhance model reliability, allow for more robust outlier detection, and better satisfy assumptions such as normality and homoscedasticity.

References

[1] 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.

[2] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. Springer.

[3] Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6), 716-723.