Modelling crime with popular instruction, attendance on religious worship and beer houses

Zoé Weissbaum

April 2024

1 Introduction

In this report we attempt to model the number of criminals per 100 000 people as a function of the number of ale and beer houses per 100 000 people, the number of attendants at school per 10 000 people and the number of attendants at public worship per 2000 people. The data was collected in England in the 1850s, and is stratified by county. Intuitively, we might anticipate that the number of criminals increases with the number of beer houses and decreases with the number of attendants at school and public worship.

The data file comprises 7 columns: `County` which contains the names of the 40 English counties, `Region` which contains the names of the regions corresponding to each county, `Code` which is a number associated to each region, `Crime` which is the number of criminals per 100 000 people for each county, `Ale` which is the number of beer houses per 100 000 people for each county, `School` which is the number of attendants at school per 10 000 people for each county and `Worship` which contains the number of attendants at public worship per 2000 people for each county. Each row within the dataset corresponds to a distinct county.

Our aim is therefore to predict the response variable `Crime` with the variables `Ale`, `School` and `Worship` through a linear regression analysis.

2 Exploratory data analysis

We first take a look at the data to get some idea of how the variables are distributed and if there is any relationship between the covariates.

We start with univariate data, meaning that we look at each variable individually and not yet at the potential relationships between them. Table 1 provides a numerical summary for each variable of interest. This gives some insight on the range and distribution of each variable. We do not notice anything specific. Figure 1 shows

	Min	Q_1	Median	Q_3	Max
Crime	66	127	157.5	174.2	241
Ale	87	209	407	490.8	708
School	560	880	965	1082.5	1250
Worship	434	654.5	801	912	1136

Table 1: Numerical summary of each variable

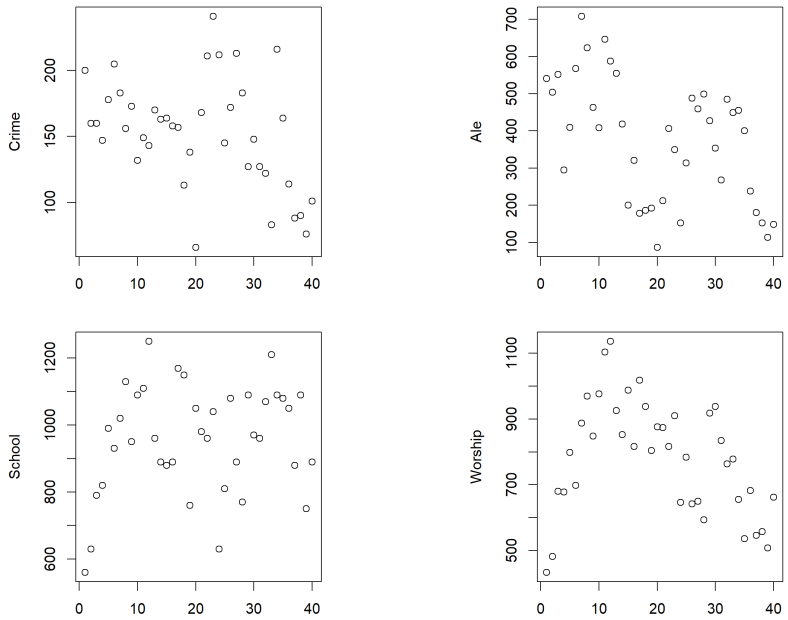

Figure 1: Scatterplots of each variable

the distribution of each variable through all 40 counties in scatterplots. Here we do not observe any particular pattern either: all variables seem to be approximately normally distributed.

We then move to bivariate data. We are now wanting to explore if there is some relation between some of the variables. On f Figure 2, we can see a strong linear association between `School` and `Worship`. These two variables are positively correlated. The same can be said about `Crime` and `Ale`. We also see these relations on table 2. We conclude from this that both `School` and `Worship` might not be necessary to predict `Crime`, and that `Ale` might explain most of `Crime`. Moreover, Table 2 indicates a slight negative correlation between `Crime` and `School`, and the absence of correlation between `Crime` and `Worship`. The negative correlation corroborates our naive assumption that the number of criminals decreases with the number of attendants at school. On the other hand, we might infer that the number of attendants at public worship has virtually no incidence on the number of criminals. This indicates that

	Crime	Ale	School	Worship
Crime	1.000	0.463	-0.230	0.004
Ale	0.463	1.000	0.135	0.153
School	-0.230	0.135	1.000	0.597
Worship	0.004	0.153	0.597	1.000

Table 2: Correlation matrix

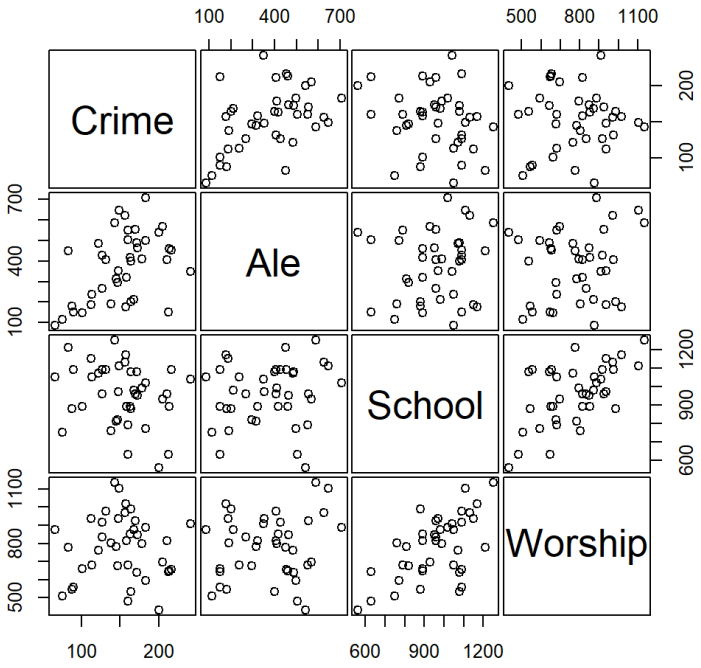

Figure 2: Bivariate relations between all variables

the variable `Worship` could possibly be left out of the model, which supports the fact that we do not need both `School` and `Worship`. We will analyse this into more detail in the model fitting.

3 Model fitting

We use linear regression to estimate the parameters of the model, meaning that we use the method of least squares.

The goal is to find an estimator $\hat{crime}$ of the response variable with estimates of the parameters $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ related to `Ale`, `School` and `Worship` respectively, as

well as an intercept $\hat{\beta}_0$ and some random error:

$$\hat{crime} = \hat{\beta}_0 + \hat{\beta}_1 \cdot ale + \hat{\beta}_2 \cdot school + \hat{\beta}_3 \cdot worship + error$$

We try fitting multiple linear models and compare them, first with adjusted R^2 and then with AIC (Aikake information criterion), to see which one fits best. The adjusted R^2 is defined as $R^2_{adj} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$, where SSE is the sum of the squares of the residuals, SST is the total sum of squares, n is the sample size and p is the number of variables in the model. The adjusted R^2 is usually preferred to the (regular) R^2 as it takes into account the number of variables. A value of R^2_{adj} closer to 1 suggests a better model. The Aikake information criterion is defined as AIC$= 2p - 2\log(\hat{L})$ where p is the number of variables and $\hat{L}$ is the maximum of likelihood. For this, a smaller value is preferred.

We first compute the simple regression models where `Crime` is explained by one variable, and compare them according to their adjusted R^2. We get

$$\hat{crime} = 109.34 + 0.12 \cdot ale + error,\ R^2_{adj} = 0.19$$

$$\hat{crime} = 209.42 - 0.06 \cdot school + error,\ R^2_{adj} = 0.03$$

$$\hat{crime} = 152.20 + 8.56 \cdot 10^{-4} \cdot worship + error,\ R^2_{adj} = -0.03.$$

This confirms our hypothesis that `Worship` is useless in predicting `Crime`. Our best model yet is the one in which `Crime` is only explained by the variable `Ale`, since it has the highest adjusted R^2 out of the three. We then use multiple regressions to fit a model with more than one covariate. We thus test the model with both `Ale` and `School` and get:

$$\hat{crime} = 178.81 + 0.13 \cdot ale - 0.08 \cdot school + error,\ R^2_{adj} = 0.26.$$

We also compare this with the full model:

$$\hat{crime} = 172.89 + 0.12 \cdot ale - 0.10 \cdot school + 0.04 \cdot worship + error,\ R^2_{adj} = 0.26.$$

These have the same value of adjusted R^2, which is higher than that of the simple `Ale` model. Since, again, the variable `Worship` does not improve the model, we will prefer the simpler model with just `Ale` and `School`.

We have found a satisfying model, but we continue our analysis further by doing a stepwise regression with AIC. We start with the full model with AIC=289.53. The AIC is most improved by removing the variable `Worship`. Now AIC=288.53. Removing other variables increases the AIC, so both remaining variables must be retained. The result is consistent with our previous findings: the best model includes both `Ale` and `School` and excludes `Worship`.

Both methods of adjusted R^2 and AIC led to the same outcome.

Our final model to estimate `Crime` is thus:

$$\hat{crime} = \hat{\beta}_0 + \hat{\beta}_1 \cdot ale + \hat{\beta}_2 \cdot school + error$$

with $\hat{\beta}_0 = 178.81$, $\hat{\beta}_1 = 0.13$ and $\hat{\beta}_2 = -0.08$ (and $\hat{\beta}_3 = 0$).

4 Model assessment

We now check that the final model fits the required assumptions and is thus valid. The model assumptions are the following: errors have mean 0, are homoscedastic (i.e. they have the same variance), are uncorrelated and are normally distributed. We see on Figure 3 that these assumptions hold approximately true. On the plot of the residuals (left), the mean of the residuals is around 0 and there is no specific pattern, but the homoscedasticity assumption seems mildly violated. The Q-Q plot (right) confirms a roughly normal distribution, even though some values show slight deviation. The standardized residuals in absolute value are smaller than 3, which suggests the absence of outliers. We conclude that, though the model does not fit the assumptions perfectly, there are no strong reasons to reject it and thus the final fitted model is suitable.

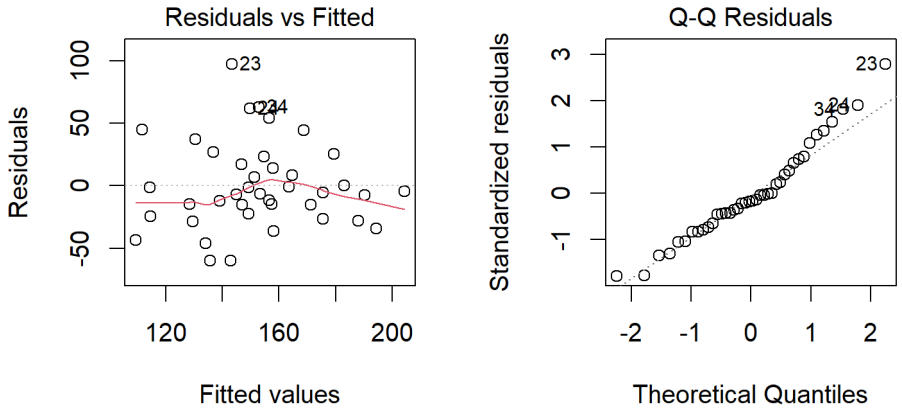

Figure 3: Residuals and Q-Q plot of the final model

5 Conclusions

In our analysis, we utilized linear regression to examine the relationship between the number of criminals per 100,000 people and the density of beer houses, school attendance rates, and attendance at public worship.

Before analysing it, we observed and summarized the data to uncover any potential patterns. The univariate summaries revealed no particular distributions, but the bivariate relations already gave some insights on how to conduct our model fitting. The absence of correlation between the variables `Worship` and `Crime` hinted that the former could be left out for the estimation of the latter.

This hypothesis was confirmed with model fitting. Through comparison of various regression models using first adjusted R^2 and then AIC, we consistently found that the crime rate `Crime` can be effectively modeled using only `Ale` and `School`. Notably, the density of beer houses emerged as the most influential predictor of crime.

This conclusion aligns with initial data exploration and echoes findings from supporting paper [1]. Specifically, we observed that the presence of beer houses correlates more strongly with increased crime rates compared to the mitigating effect of school attendance and public worship attendance. Additionally, while school attendance and attendance at public worship are correlated, the former exerts a more significant influence on mitigating criminal activity.

6 Bibliography

[1] J. Clay, "On the relation between crime, popular instruction, attendance on religious worship, and beer-houses," *Journal of the Statistical Society of London*, Vol. 20, 1857.

[2] D. G. Rossiter, "An example of statistical data analysis using the R environment for statistical computing," 2014.