Assignment 2: Moral Statistics Revisited

INTRODUCTION TO STATISTICS FOR DATA SCIENCE

Fergus Walsh

Due 18th January, Epiphany Term, 2021

> "We are forced to recognise that in
> many respects judicial statistics
> represent complete certainty."
>
> André-Michel Guerry
> Letter to Adolphe Quételet
> 11 September 1831

1 Introduction

1.1 Guerry: His Life, Times and Work

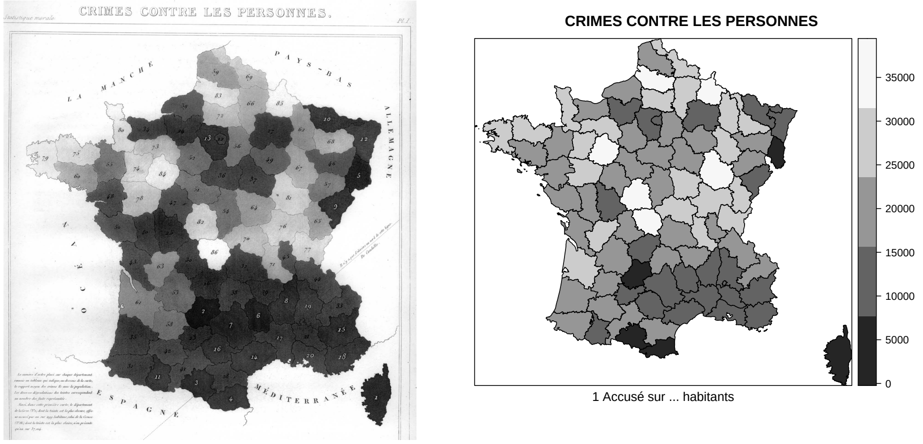
André-Michel Guerry (1802–1866) was a pioneering figure in the fields of sociology and crimionology, and was one of the first to systematically compile, analyse and present data on crime.

In the 1820's the French government began to regularly collect data on crime reporting, conscription, taxation, wealth and various other social aspects of the population. For the first time, data was regularly available on multiple social and criminal phenomena on a national scale, and from 1827 published annually in the *Compte Général de l'Administration de la Justice Criminelle en France*. Working as a lawyer at the Ministry of Justice, Guerry devoted his life to understanding this data and publicising his findings. As he notes himself, "Today, theoretical systems do not remain abstractions, but move from their written form into the public sphere, and then into our institutions, where they exert either a healthy or damaging influence on society."[1] This interest in criminal statistics and their communication led Guerry to publish his *Essai sur la statistique morale de la France* in 1833, a detailed presentation of criminal and societal statistics and analysis of the relationships in the data. Guerry supplemented his analysis with choropleth maps

[1] All translations my own. Guerry 1833, pp. i–ii; Friendly 2007, pp. 369–370; Hacking 1990, pp. 73–77.

(Figure 1), the popularisation of which Guerry was most famous for in his own day and on which his legacy rests.[2]

Guerry's method in part rested on the comparison of choropleth maps depicting different variables, which allowed the reader to directly compare the differing rates of crime, suicide, literacy, *et cetera* between the different départments of France. Although this technique seems crude to us today, Guerry was active before the widespread adoption of the scatterplot and regression.[3]



(a) Plate I, Guerry 1833

(b) Plate I reproduced in R

FIGURE 1: Choropleth Map showing Population per Crime against Persons, Guerry's 1833 original and the same data plotted in R with the `spplot()` function. In both cases the darker shading indicates a worse crime-rate. Guerry's plot shows the departments ranked by crime-rate, rather than the values themsleves.

1.2 *Essai sur la statistique morale de la France* Revisited

What if, however, Guerry had used modern statistical techniques in his analysis, would he have arrived at the same conclusions ? As with Guerry's 1833 *Essai*, this enquiry will research the patterns between societal statistics and crime rates, and attempt to model the relationships using least squares linear regression. In his *Essai*, Guerry considered crime against persons (such as homicides, assualts, sexual offences but also bigamy and perjury), crime against property (such as theft, robbery (including violent robbery), fraud, tax evasion) and suicide, but this enquiry will be limited only to crime against persons for reasons of brevity.[4]

[2] In 1864 the Academy of Sciences honoured Guerry with an award, "not for the facts but for their display, their marvellous maps of crime and suicide." HACKING 1990, p. 76.

[3] FRIENDLY 2007, p. 378.

[4] I must, of course, note the similar analyses of FRIENDLY 2007 and WHITT 2010 who both deal with Guerry's data in far more detail than offered in this enquiry.

1.3 Overview of the Data

The data sets used in this enquiry come from the 'Guerry' package for R, prepared by Michael Friendly and Stephane Dray.[5] The package contains the data used by Guerry in the *Essai* (Guerry) and Adolph d'Angeville in his 1836 *Essai sur la Statistique de la Population française* (Angeville), along with maps of France following the national and départment boundaries of 1830 (gfrance and gfrance85). Both Guerry and d'Angeville drew on the *Compte Général*, while d'Angeville also drew on the census of 1831. I have therefore used both data sets, as in some cases Angeville provides better information than the Guerry data set. The values given for most variables are often the mean figures per départment between 1825 or 1827 and 1830, depending on the availability of the data.[6]

Table 1 details the 23 variables in the Guerry data set, while Table 2 details the 16 variables in the Angeville data set. Guerry standardised crime, suicide and certain other statistics as population per variable (popuation ÷ no. crimes), rather than variable per capita as we would today (no. crimes ÷ population). Thus crime against persons (Crime_pers) is given as the population per each accusation of a crime against the person: there are only 2,199 inhabitants per accusation in Corsica, whereas in Creuse there are 37,014 inhabitants per accusation. However, certain other variables (e.g. Commerce, Clergy) and most of the Angeville data is recorded as per capita figures. In addition, in the cases of Legit_births, Illeg_births, Recruits, Farmers and Primary_schools, the Angeville data set records the absolute values from the 1831 census, and so these values were divided by the population of each départment to give a standardised value. Whenever these variables are discussed in the modelling process, they were used in this per capita form. A population density statistic was also created for each départment by dividing Pop1831 by Area.

Finally, a number of variables in Guerry are presented as rank measures, ordering the départments from 1st to 86th, with the order always arranged from 'best' to 'worst'. Therefore Corsica (1) has the highest number of Catholic priests per captia, and Charente-Inferieure (86) has the fewest; whereas Vienne (1) has the smallest ratio of infanticides per capita, and Oise (86) has the greatest. Ranking the départments in this way was necessary to produce the choropleth maps, as a darker shade always indicated a worse result in every variable.[7]

Although the *Compte Général* purported to record figures accurately, Guerry clearly had some reservations as to the accuracy of the data. This is unsuprising, given that only after Guerry's own instigation did the Ministry of Justice begin to record more complete suicide data, including, "the easiest facts, namely the age and sex of suicides."[8] "Instead of the number of convictions," Guerry wrote, ". . . we have used the number of persons accused, which more accurately represents the number of crimes commited." He later argued, "for arson, more than three-quarters of the accused are acquited each year, even though there is almost never any doubt as

[5] Friendly et Dray 2020.

[6] Documentation, Friendly and Dray 2020, pp. 8–9.

[7] Friendly 2007, p. 373.

[8] Hacking 1990, p. 79.

to the reality of the crime. In a majority of départments, infanticide remains the crime least often punished, whereas in others its repression is incredibly severe."[9] Guerry's own solution was to examine statistics by larger regions, though that led him into the ecological fallacy.[10] In addition, by converting the absolute values into ordinal ranks for each département, extreme variations in the data may have also been mitigated, since all observations are of equal steps apart from each other, and therefore extreme outliers are not possible.

[9] Guerry 1833, pp. 6–7.

[10] Guerry 1833, pp. 7, 9; Whitt 2010, p. 134.

Variable	Definition	Notes
dept	Department numerical I.D.	1–19, 21–89, Corse (Corsica) as 200. Corsica was No. 20, but later it was split into 20A and 20B.
Region	Region of France	North(N), South (S), East (E), West (W), Central (C) and Corsica (NA).
Department	Department name	As in 1830, without diacritics.
Crime_pers	Population per crime against persons	Guerry 1833, pp. 38–41
Crime_prop	Population per crime against property	Guerry 1833, pp. 42–44
Literacy	% of department able to read and write	Guerry 1833, pp. 45–51
Donations	Population per no. of donations to the poor	Frequency of donation per capita, rather than the value. Guerry 1833, pp. 56–58
Infants	Population per illegitimate birth	Guerry 1833, pp. 52–55
Suicides	Population per suicide	Guerry 1833, pp. 61–69
MainCity	Size of principal town in each department	Small (1:Sm) for the smallest ten cities, Large (3:Lg) for the largest ten cities, all others Medium (2:Med).
Wealth	Per capita property tax	Rank measure: 1 = most wealthy départment. Guerry 1833, p. 70
Commerce	No. patents per capita	Rank measure: 1 = most industrious départment. Used to measure commerce and industry. Guerry 1833, p. 70
Clergy	No. active Catholic priests per capita	Rank measure: 1 = greatest no. priests. Guerry 1833, p. 70
Crime_parents	Ratio of crimes against parents : all crimes	Rank measure: 1 = lowest proportion of crimes against parents. Guerry 1833, p. 70
Infanticide	Infanticides per capita	Rank measure: 1 = lowest proportion of infanticides. Guerry 1833, p. 70
Donation_clergy	Donations to religious establishments per capita	Rank measure: 1 = largest no. of donations per capita. Guerry 1833, pp. 58–59
Lottery	Wager on Royal Lottery per capita	1 = greatest proceeds in lottery. Guerry 1833, p. 70
Desertion	Military desertion	Rank measure of ratio of deserters : size of military unit: 1 = fewest deserters. Guerry 1833, p. 70
Instruction	Level of education	Rank measure: 1 = highest level of education, based on literacy test given to army recruits. Guerry 1833, pp. 45–46
Prostitutes	No. prostitutes in Paris	Classified by department of birth.
Distance	Distance of each department to Paris (km)	Calculated from centroids in gfrance spatial polygons data frame object. Documentation, Friendly and Dray 2020, p. 9
Area	Area of department (1,000km^2)	From d'Angeville 1836
Pop1831	Population of department (1,000s)	From d'Angeville 1836

Table 1: Variables in the Guerry dataset

Variable	Definition & Notes
dept	Département numerical I.D. = Guerry dept
Department	Département name = Guerry Department
Mortality	No. births per 100 21-year-old inhabitants
Marriages	No. marriages per 1,000 21-year-old men
Legit_births	No. legitimate births per annum
Illeg_births	No. illegitimate births per annum
Recruits	No. people registered for military recruitment
Conscripts	No. inhabitants per military conscript
Exemptions	No. exemptions from military service per 1,000 due to physical causes
Farmers	No. farmers, absolute value from 1831 census
Recruits_ignorant	Mean no. ignorant (i.e. illiterate) recruits per 1,000
Schoolchildren	No. schoolchildren per 1,000
Windows_doors	No. windows and doors in houses per 100 inhabitants. Serves as an indicator of household wealth
Primary_schools	No. primary schools
Life_exp	Life expectancy (years)
Pop1831	Population of department (1,000s) = Guerry Pop1831

Table 2: Variables in the Angeville dataset. All details from Documentation Friendly and Dray 2020, pp. 4–5

1.4 Preliminary Correlations

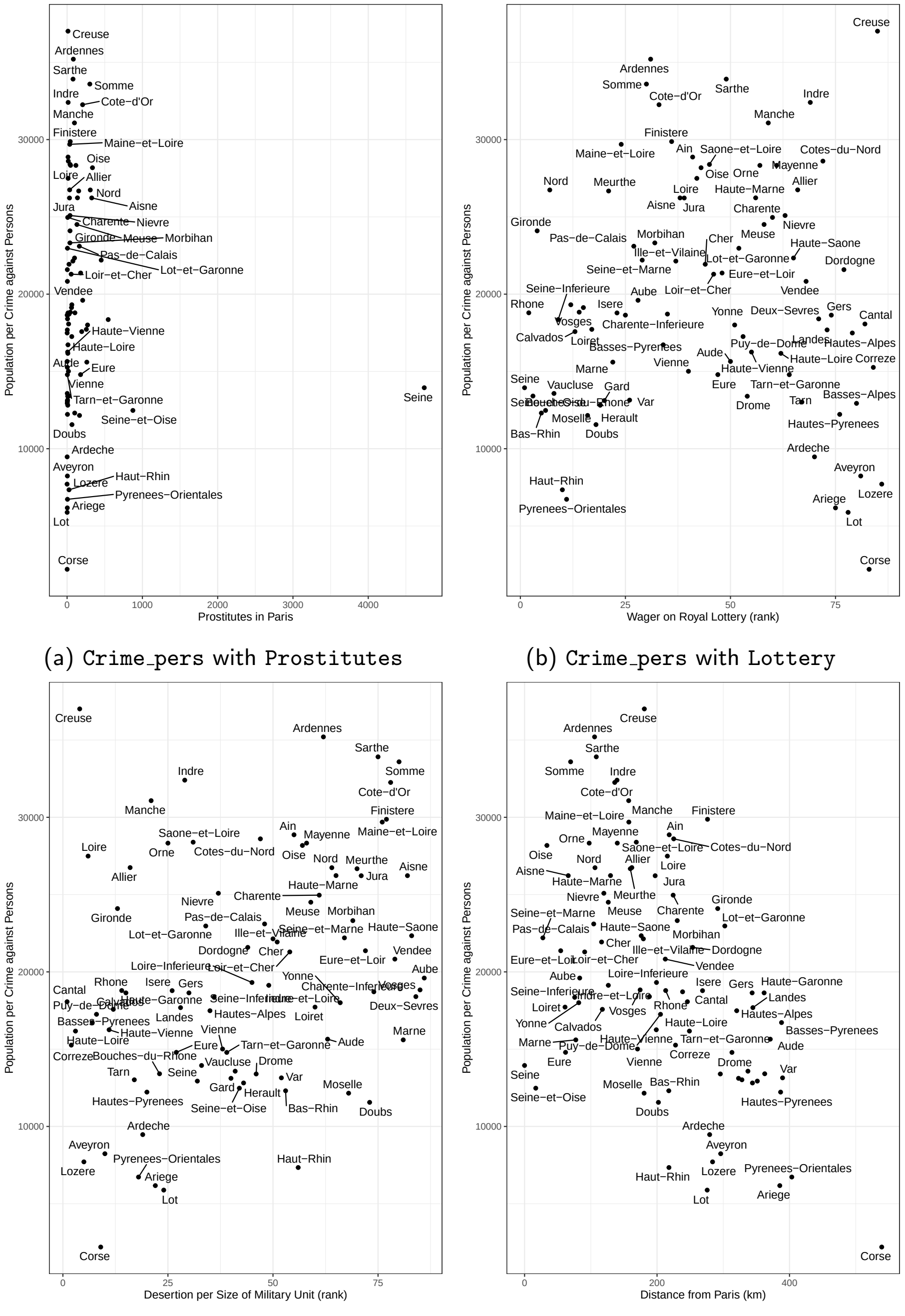
Table 3 shows the bivariate correlations between Crime_pers and selected other variables. The majority of coefficent scores of variables correlated with Crime_pers is close to 0. Where there is correlation, the coefficent is not very high; the maximum being –0.509 (r) and –0.489 (r_s) in the case of Crime_pers and Distance. At first glance, there also appear to be two non-linear correlations (Prostitutes and Lottery), which are plotted in Figure 2.

Desipte the r_s figure of 0.411, Prostitutes and Crime_pers do not have a real correlation, even when the outlier Seine is removed. Rather, since Prostitutes only records the number of prostitutes in the city of Paris, rather than the whole of France, this bivariate relationship can be considered an extreme example of a funnel plot. The different values are what one would expect from a random variable of this kind and are best explained by chance alone, rather than there being any link between a départment's crime rate and the number of Parisian prostitutes who were born there.[11]

Despite the r_s value of 0.008, Lottery does in fact appear to have a negative quadratic correlation with Crime_pers (Figure 2b). When the outlier Creuse is removed, the r_s figure increases to –0.026.

As can be seen from Table 3 and Figure 2, the départments of Seine, Corse and

[11] This is analogous to Spiegelhalter's example of differing bowel cancer rates in U.K. health boards, "Smaller districts have fewer cases and so are more vunerable to the role of chance, and therefore tend to have more extreme results." Spiegelhalter 2019, pp. 233–236.



(a) Crime_pers with Prostitutes

(b) Crime_pers with Lottery

(c) Crime_pers with Desertion

(d) Crime_pers with Distance

Figure 2: Scatter plots showing correlation between Crime_pers and selected other variables

Variable	r	r_s	Outliers
Desertion	0.331	0.342	Creuse
Crime_prop	0.275	0.233	
Infanticide	0.275	0.271	Corse
Prostitutes	0.069	0.411	Seine, Seine-et-Oise
Lottery	0.003	0.008	Creuse
Suicides	−0.130	-0.055	Aveyron, Ariege, Hautes-Pyrenees, Haute-Loire
Donation_clergy	−0.180	−0.177	Creuse
Crime_parents	−0.195	−0.193	Creuse, Indre
Distance	−0.509	−0.489	

Table 3: Pearson's correlation coefficient (r), Spearman's rho correlation coefficent (r_s) and outliers for the correlation between Crime_pers and selected other variables, given to 3 decimal places.

Creuse are often outliers. Seine (and the nearby Seine-et-Oise) covers the city of Paris, which has the greatest population density by far and the urban character of this département explains its outlier status in variables such as Mortality, for instance.[12] Corsica (Corse), visible as outlier in Figure 2, was an almost lawless island in the 18th and early 19th centuries; where *vendettas* and banditry were common, hence its high crime rate (expressed as a small value in Guerry's scale).[13].

2 Modelling

2.1 Method

To examine the possible links between crime and the available social data, crime against persons (Crime_pers) is modelled using a multiple least squares regression. The model was constructed by testing statistical significance of variables using the Student t-test, by comparing the adjusted R^2, Akaike information criterion (A.I.C.) and Bayesian information criterion (B.I.C.) of a number of candidate models, and by k-fold cross validation.

This is the process whereby the data set was randomly split into five sections ($k = 5$), each model was trained on four sections, and then each model's prediction was compared to the fifth unused section, this is then repeated k times.[14] Each time the mean squared error (M.S.E.) statistic is calculated, being the error between the model's prediction and the part of the data left out. The mean of $MSE_1, MSE_2, MSE_3, MSE_4$ and MSE_5 is the cross validation statistic (C.V.) and

[12] Calculated as Pop1831 ÷ Area. Seine has a value of 1.227, the next densest département Nord only 0.172

[13] Guerry 1833, p. 37; Zamoyski 2018, pp. 9–13; Whitt 2010, p. 131.

[14] Note, due to having NA values for Recruits_ignorant and Windows_doors, Corsica was not included in the cross-validation data set, but is included in the diagnostics for the final model (section 2.5)

is used as a measure of the model's predictive performance.[15] This process was carried out using the `cv.glm()` function in R, which automatically divides the data, carries out the validation and calculates the C.V. value.[16]

2.2 Modelling a Non-Linear Relationship

Earlier, it was noted that there was a negative quadratic correlation between `Crime_pers` and `Lottery`. By modeling this relationship between these two variables using linear regression, the hypothesis that these two variables are related quadratically can be tested. Namely, that for the simple model,

$$Y = \beta_0 + \beta_1 \times \texttt{Lottery} + \varepsilon \tag{1}$$

firstly, that

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

and secondly, that when modelled with a polynomial as,

$$Y = \beta_0 + \beta_1 \times \texttt{Lottery} + \beta_2 \times \texttt{Lottery}^2 + \varepsilon \tag{2}$$

the hypothesis test is:

$$H_0 : \beta_2 = 0$$
$$H_1 : \beta_2 \neq 0$$

In addition, using analysis of variance techniques, the hypothesis that the fit of the polynomial model is better than the fit of simple linear model can be tested. That is,

$$H_0 : \beta_1 = \beta_2 = 0$$
$$H_1 : \text{ at least } \beta_1 \text{ or } \beta_2 = 0$$

Table 4 shows the summary statistics for the linear model, trained on the whole data set. As the *p*-value for the coefficent of `Lottery` is relatively high, the null hypothesis cannot be rejected. However, in the case of the polynomial model (Table 5), the *p*-values for both `Lottery` and `Lottery`^{2} are below the 0.05 level, and so the null hypothesis can be rejected.

	Coefficent	Standard Error	*t*-statistic	*p*-value
Intercept	19,720.000	1,642.000	12.007	< 0.001
`Lottery`	0.783	32.790	0.024	0.981

Table 4: Summary statistics for the model $Y = \beta_0 + \beta_1 \times \texttt{Lottery} + \varepsilon$, given to 3 d.p.

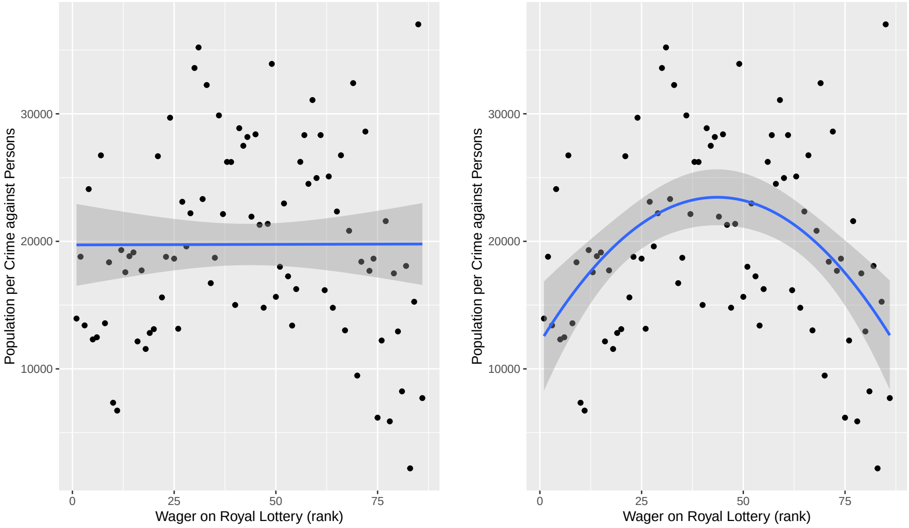
[15] James et al. 2013, pp. 175–176.

[16] 'boot' package documentation, Ripley 2020, pp. 41–43.

	Coefficent	Standard Error	t-statistic	p-value
Intercept	12,056.558	2254.058	5.349	< 0.001
`Lottery`	523.314	119.582	4.376	< 0.001
`Lottery`^{2}	–6.006	1.332	–4.510	< 0.001

Table 5: Summary statistics for the model $Y = \beta_0 + \beta_1 \times \texttt{Lottery} + \beta_2 \times \texttt{Lottery}^2 + \varepsilon$, given to 3 d.p.

Table 6 shows the analysis of variance statistics computed by comparing the two models with the `anova()` function in R. The numer of observations in the data sets is 86, a large n value (> 30); therefore, since the F-statistic is 20.337 and the associated p-value is close to 0, the null hypothesis can be rejected.[17]



(a) Linear Model

(b) Polynomial Model

Figure 3: Linear and Polynomial Models of `Lottery`. The blue line in each case shows the predicted values of the two models, with the grey shading indicating the 95% confidence interval.

It seems to be the case that `Lottery` should be modelled with a polynomial. This hypothesis can now be validated in turn by testing a linear and polynomial model each trained on a k-fold of the data, and comparing the cross-validation statistic (Table 7).

The polynomial model has a smaller R.S.S. than the linear model, and hence the variable `Lottery` will be modeled using a polynomial function. However, the

[17] "When n is large, an F-statistic that is just a little larger than 1 might still provide evidence against H_0. In contrast, a larger F-statistic is needed to reject H_0 if n is small." James et al. 2013, pp. 75–76, 116.

	Res. Deg. Freedom	Residual Sum of Squares	F-statistic	p-value
`Lottery`	84	4,787,215,801	—	—
`Lottery`^{2}	83	3,845,084,557	20.337	< 0.001

Table 6: Analysis of Variance table comparing $Y = \beta_0 + \beta_1 \times \texttt{Lottery} + \varepsilon$ and $Y = \beta_0 + \beta_1 \times \texttt{Lottery} + \beta_2 \times \texttt{Lottery}^2 + \varepsilon$ given to 3 d.p.

	Cross-Validation Statistic
`Lottery`	58,722,508
`Lottery`^{2}	47,381,367

Table 7: Cross-Validation Statistics from the `cv.glm()$delta` Function in R

cross-validation values are quite large. Looking again at Figure 3b, this is not too suprising given the high variance of the error terms.

2.3 Modelling `MainCity` using Dummy Variables

`MainCity` records whether the principal town of the département is the ten largest (`3:Lg`), ten smallest (`1:Sm`) or between these two extremes (`2:Med`). To model this, R generated two dummy variables for `MainCity` when generating a linear regression model with `MainCity` as a predictor variable: `MainCitySm` if the value is `1:Sm`, `MainCityLg` if the value is `3:Lg` and neither variable if the value is `2:Med`.[18] In the R console, this is displayed as:

```
Call:
lm(formula = Guerry$Crime_pers ~ Guerry$MainCity, data = Guerry)
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         20352.2      924.9  22.005   <2e-16 ***
Guerry$MainCitySm  -2564.4     2549.7  -1.006    0.317
Guerry$MainCityLg  -2576.6     2549.7  -1.011    0.315
> contrasts(MainCity)
      Sm Lg
Med    0  0
Sm     1  0
Lg     0  1
```

which corresponds to the model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon & \text{if city is } \texttt{1:Sm} \\ \beta_0 + \beta_2 + \varepsilon & \text{if city is } \texttt{3:Lg} \\ \beta_0 + \varepsilon & \text{if city is } \texttt{2:Med} \end{cases} \tag{3}$$

[18] Note, for R to automatically convert `MainCity` into dummy variables, the values were manually rewritten as `Sm`, `Med` and `Lg` without the numbers.

As can be seen, the two variables have p-values which do not suggest that MainCity is a statistically significant variable. Rather, an F-test should be used to test the hypothesis,

$$H_0 : \beta_1 = \beta_2 = 0$$

$$H_1 : \beta_1 \neq \beta_2 \neq 0$$

Table 8 shows the Analysis of Variance table for this model. The F-statistic is very low and below the 95% critical value, displayed by the high p-value. The null hypothesis cannot therefore be rejected, and it must be concluded that the character of a département's principal town, whether it be very urban or very small, cannot be used to predict the corresponding rate of crimes against persons. This variable, therefore, will not be used in the linear regression model.[19]

	Degrees of Freedom	Sum of Squares	F-statistic	p-value
MainCity	2	101,417,490	0.898	0.411

Table 8: Analysis of Variance table for Crime_pers modelled with MainCity using two dummy variables, given to 3 d.p.

2.4 Variable Selection & Cross-Validation

Having dealt with MainCity and Lottery, it is now necessary to select which other variables should be included in the model. The model will be refined by examining the statistical significance of the coefficents of the variables (t-test), the fit of the model onto the data (adjusted R^2) and by cross-validating the model using k-fold resampling (M.S.E.).

Let **Model A** include all variables, with the exception of MainCity, dept, Department and Region, with Legit_births, Illeg_births, Recruits, Farmers and Primary_schools in their standardised form, with Lottery as a polynomial and with the population density statistic (Pop1831 ÷ Area).[20]

Only 6 of the variables in Model A are statistically significant; **Model B** was generated by reducing the number of variables using the backwards and step-wise elimination function in R. Model B has a slightly better fit than Model A, the adjusted R^2 and A.I.C. value are slightly improved, and the average M.S.E. has decreased by about a third. As can be seen in Table 9, however, there are only 8 significant variables in Model B with a critical value at the 0.05 level: Infanticide, Lottery, Distance, Area, Marriages, Conscripts, Farmers and Primary_schools. These 8 variables were taken to generate **Model C**, detailed in Table 10.

Model C has worse adjusted R^2, A.I.C. and R.S.E. values than Model B, though this is to be expected as the number of variables has been halved. However, the cross-validation statistic is about half of Model B's and a third of Model A's, while

[19] In addition, the population density statsitc (Pop1831 ÷ Area) also turned out to be an insignificant variable.

[20] Region is not included as it could lead to a model based on the ecological fallacy.

	Coefficent	Standard Error	*t*-statistic	*p*-value
`Intercept`	91,265.859	45,339.958	2.013	0.048
`Literacy`	135.767	74.064	1.833	0.071
`Infants`	−0.165	0.103	−1.610	0.112
`Commerce`	69.720	32.697	2.132	0.037
`Infanticide`	94.418	25.715	3.672	$<$ 0.001
`Lottery`	315.524	126.461	2.495	0.015
`Lottery`^{2}	−2.722	1.336	−2.038	0.046
`Desertion`	46.725	29.695	1.574	0.121
`Prostitutes`	−5.225	3.030	−1.724	0.089
`Distance`	−36.714	8.547	−4.295	$<$ 0.001
`Area`	1.126	0.466	2.418	0.018
`Mortality`	177.936	130.136	1.367	0.176
`Marriages`	29.288	10.261	2.854	0.006
`Legit_births`	−2,199.942	1,405.922	−1.565	0.123
`Illeg_births`	−2,835.850	1,926.047	−1.472	0.146
`Recruits`	1,134.119	870.386	1.303	0.197
`Conscripts`	−798.886	388.043	−2.059	0.044
`Exemptions`	−4.360	2.982	−1.462	0.149
`Farmers`	−15.224	5.985	−2.544	0.013
`Windows_doors`	−63.970	37.921	−1.687	0.096
`Primary_schools`	−6,672.095	2,915.599	−2.288	0.025

Table 9: Summary statistics for Model B, given to 3 d.p.

the *F*-statistic has likewise improved. Model C, therefore, has a much looser fit to the training data, but performs far better in the cross-validation testing: Models A and B represent both the signal and the noise, whereas Model C is far more focused on the signal only.

The focus now becomes on fine-tuning the model's complexity to find the optimum trade-off between bias and variance. **Model D** is Model C without `Marriages`, the variable with the lowest *t*-statistic and highest *p*-value; **Model E** is Model C without `Farmers`, the second least insiginifcant variable in Model C, while **Model F** contains neither.[21] Finally, **Model G** is the same as Model F but without `Conscripts`, Model F's least significant variable.[22]

As more variables are removed from Model C, the cross-validation statistic no longer steadily falls. Indeed, the values for Models D-G vary each time the `cv.glm()$delta` function is run in R, and although it would have been more rig-

[21] In R notation, D: `lm(Crime_pers ~ Infanticide + Lottery + I(Lottery^2) + Distance + Area + Conscripts + Farmers + Primary_schools, data = Guerry)`
E: `lm(Crime_pers ~ Infanticide + Lottery + I(Lottery^2) + Distance + Area + Conscripts + Marraiges + Primary_schools, data = Guerry)`
F: `lm(Crime_pers ~ Infanticide + Lottery + I(Lottery^2) + Distance + Area + Conscripts + Primary_schools, data = Guerry)`

[22] `lm(Crime_pers ~ Infanticide + Lottery + I(Lottery^2) + Distance + Area + Primary_schools, data = Guerry)`

	Coefficent	Standard Error	t-statistic	p-value
Intercept	43,336.150	11,329.687	3.825	< 0.001
Infanticide	109.307	26.296	4.157	< 0.001
Lottery	389.252	109.405	3.558	0.001
Lottery^2	–3.846	1.198	–3.211	0.002
Distance	–38.043	6.787	–5.605	< 0.001
Area	1.050	0.455	2.307	0.024
Marriages	2.234	7.704	0.290	0.776
Conscripts	–234.896	100.842	–2.329	0.026
Farmers	–9.047	5.047	–1.793	0.077
Primary_schools	–4,546.784	1,581.302	–2.875	0.005

Table 10: Summary statistics for Model C, given to 3 d.p.

Model	p	Adj. R^2	A.I.C.	R.S.E.	F-statistic	p-value	C.V.
A	32	0.532	1,697.59	4,994	4.079	< 0.001	96,108,428
B	21	0.590	1,717.78	4,675	7.039	< 0.001	64,508,305
C	10	0.508	1,707.80	5,263	10.760	< 0.001	34,063,628
D	9	0.542	1,711.90	5,338	11.370	< 0.001	31,284,996
E	9	0.514	1,706.00	5,232	12.240	< 0.001	33,136,032
F	8	0.498	1,710.40	5,320	13.020	< 0.001	32,103,978
G	7	0.468	1,714.00	5,476	13.440	< 0.001	32,851,115

Table 11: Summary statistics given to 3 d.p.

orous to run the function a number of times and take the mean C.V. statistic, this turns out not to be necessary. For the purposes of selecting the optimum model, since the negative change in the adjusted R^2 and the positive change in the R.S.E. values between Models E-G indicate that these models are not an improvement over Model D, Model D represents the best trade-off between model bias and variance. It performs as well as Models E-G in cross-validation testing, but also has a slighlty closer fit to the training data, and is the model ultimately selected.

2.5 Model Diagnostics & Assumptions

In *An Introduction to Statistical Learning*, the authors identify five potential problems that can arise from the assumptions of the least squares linear regression, namely:[23]

1. Non-linearity of the response-predictor relationships.

2. Correlation of error terms.

3. Non-constant variance of error terms.

4. Outliers.

[23] James et al. 2013, pp. 92–101.

	Coefficent	Standard Error	t-statistic	p-value
Intercept	43,181.731	11,172.582	3.865	< 0.001
`Infanticide`	97.989	26.090	3.756	< 0.001
`Lottery`	409.205	108.316	3.778	< 0.001
`Lottery`^{2}	−3.953	1.180	−3.351	0.001
`Distance`	−31.632	7.999	−3.954	< 0.001
`Area`	1.208	0.462	2.618	0.012
`Conscripts`	−208.325	72.739	−2.864	0.005
`Farmers`	−15.129	6.211	−2.436	0.017
`Primary_schools`	−4,571.103	1,506.066	−3.035	0.003

Table 12: Summary statistics for Model D, given to 3 d.p.

5. High-leverage points.

6. Collinearity.

The issue of non-linear relationships was dealt with in section 2.2. A plot of the residual and the predicted values (Figure 4a) shows that Model D adequately satisfies this assumption. Although there is a slight wave pattern to the trend line, it appears to vary equally in either direction.

The second assumption is that the error terms are unrelated, i.e. that ε_1 gives no information about ε_2 or any other error term. From Figure 4a, this seems to be the case, and if so the Spearman rank correlation test value should be 0. This is equivalent to the hypothesis test:

$$H_0 : r_s = 0$$

(The error terms are unrelated.)

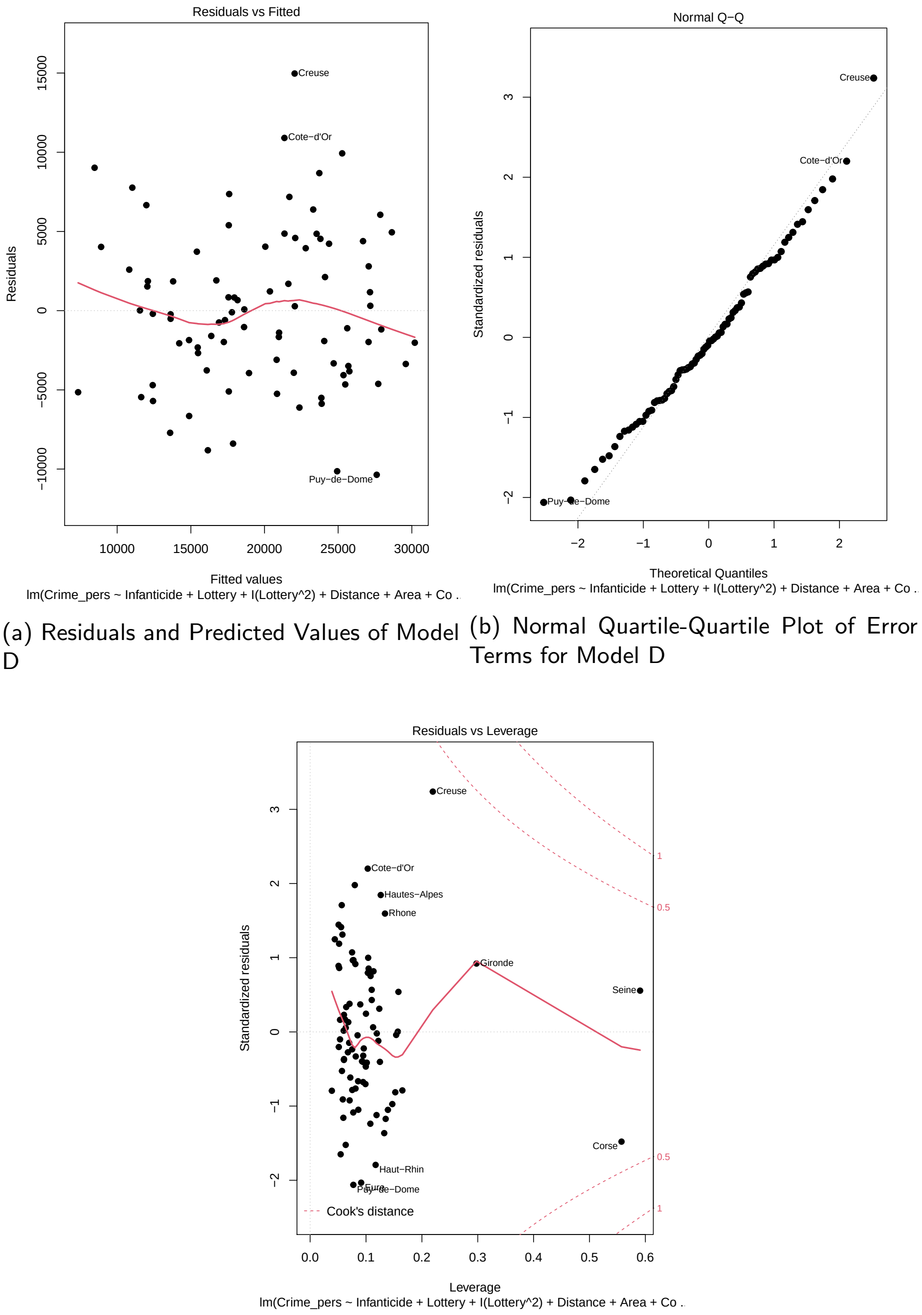
$$H_1 : r_s \neq 0$$

(The error terms are related.)

The r_s figure is 0.0004885675 and the p-value is 0.997 (3 d.p.), meaning that the null hypothesis cannot be rejected, with almost certainly. Moreover, even if it were rejected, the r_s value would be tiny.

Likewise, it is assumed that the error terms have constant variance, that is $Var(\varepsilon_i) = \sigma^2$ Again, Figure 4a seems to suggest that there is homoscedasticity, as there is no obvious increase in variance as the predicted values increase. Furthermore, a normal quartile-quartile plot (Figure 4b) also shows that the variance of the error terms is normally distributed (with the exceptions of a handful of extremes including `Creuse`, `Eure` and `Puy-de-Dome`). This assumption will therefore be taken to be satisfied.

Outlying observations can adversely affect the R.S.E. and R^2 values for a model. Outliers in certain instances were noted in Table 3 and treated in section 1.4. As noted above, the variance of the residuals is reasonably equal in both the positive



(a) Residuals and Predicted Values of Model D

(b) Normal Quartile-Quartile Plot of Error Terms for Model D

(c) Residuals and Leverage Statistics of Model D

Figure 4: Diagnostic Plots for Model D

and negative directions, and so it is unlikely that any outlier will be adversely affecting the model. This can be confirmed by calculating the studentised residuals (`rstudent()` in R). For Model D, these range between –2.107 (`Puy-de-Dome`) and 3.462 (`Creuse`). This is not an excessive range, and as such this assumption can be taken to be satisfied.[24]

	Département	Leverage Statistic
Maximum	`Seine`	0.591
	`Corse`	0.557
	`Gironde`	0.298
	`Creuse`	0.220
Mean		0.116
Minimum	`Loir-et-Cher`	0.039

Table 13: Leverage statistics for certain observations and the mean leverage statistic for Model D, given to 3 d.p.

Similarly, observations with high leverage can also adversely affect the regression line. Figure 4c shows the leverage statistic (h_i) plotted against the residuals and Table 13 details certain values of the leverage statistics.[25] Four observations have very high h_i values, yet I do not believe it is necessary to remove these four observations from the model. In their modelling of the Guerry and d'Angeville's data, Whitt removed Corsica from their modelling process, "for technical reasons [involivng software, and] . . . since it was an extreme outlier, by far the most violent and least developed of all départments. Though technically part of France, it more closely represented a foreign colony."[26] Nevertheless, although one should be aware of their high leverage, since both `Corse` and `Seine` are not also outliers in the residual and fitted plot, I do not believe it is not necessary to remove them from the model.

Finally, it is assumed that no two predictor variables are closely related to each other, termed collinearity. The extent of collinearity in a model can be tested by calculating the variance inflation factor (V.I.F.) for each coefficent. As Table 14 shows, only `Lottery` and `Lottery`^{2} have high V.I.F. values.[27] Yet this is to be expected, the two variables are in fact closely related to each other.

Model D, therefore, satisfies the assumptions of the linear regression modelling process.

[24] "The outlier has a studentized residual of 6; typically we expect values between –3 and 3." James et al. 2013, caption to Figure 3.12, p. 97.

[25] Note, the mean value is given by the formula $\frac{p+1}{n}$ where p is the number of coefficents. Here it is therefore $\frac{10}{86}$ rather than 0.105 as given by `mean(hatvalues())`. James et al. 2013, p. 98

[26] Whitt 2010, p. 150.

[27] "a V.I.F. value that exceeds 5 or 10 indicates a problematic amount of collinearity." James et al. 2013, pp. 101–102.

Coefficent	Variance Inflation Factor
Infanticide	1.255
Lottery	22.798
Lottery^2	22.018
Distance	1.674
Area	1.244
Conscripts	2.074
Farmers	1.846
Primary Schools	1.240

Table 14: Variation inflation factor statistics for Model D, given to 3 d.p. Note the intercept β_0 does not have a V.I.F. value.

3 Discussion of Results

3.1 Findings

The most surprising variable in the model may be that distance from Paris is a negative predictor (–38) and that it had the strongest correlation (r –0.509) with crimes against persons. Whitt treats Distance as an effective indicator of state control, which decreases steadily outside of Paris and away from the major transportations routes between Paris and the large cities.[28] At first glance, the negative correlation might put the efficacy of the collection and compiling of the data in doubt (i.e. away from Paris the administration is less rigorous in recording crimes), yet Guerry recorded population per crime, and so a smaller value indicates a higher crime rate. One should also note the negative coefficent of the predictor Farmers which indicates that the rate crimes against persons was greater in more agrarian départments. These relationships in the model point to a larger historical and sociological hypothesis, that in industrialised regions suicide is more common that homicide, whereas in rural regions the reverse is true.[29] This idea is not the subject of this equiry, but it should be noted that the notion of an inverse relationship between homicide and suicide began with Guerry and his contemporaries, and continues to excite discussion in sociology to this day.

One of Guerry's main aims in the *Essai* was to test the hypothesis that a greater level of education resulted in a lower crime rate.[30] Indeed, both Literacy and Instruction were found to be insignificant variables, but Primary_schools has the largest coefficent in the model, –4,571. Rather, I propose that this variable does not only indicate the provision of education, but also serves as an indicator of the demographic distribution of the population, such as the birth rate or number of children. The model would imply that these two statistics are linked, which can be corroberated by the inclusion of Infanticides in the model.[31] Guerry

[28] Whitt 2010, p. 155.

[29] For an overview of the literature, Whitt 2010, pp. 130–132.

[30] "Ignorace, people say, is the primary cause of crime ... our map provides evidence to the contary. ... Evidently the link which people talk about does not exist. Guerry 1833, pp. 45–46, 51.

[31] Note due to the rank scale of Infanticides, a positive coefficent indicates an increase in the

recorded that infanticides were the seventh most frequent category of crime against persons (63 per 1,000 crimes) and the most frequent category of crime committed by women (406 per 1,000).[32] Given these proportions, it is perhaps not surprising that Infanticides is a significant predictor in the model.

Guerry did not treat Conscripts as a significant variable in his analysis, but it is not suprising that greater rates of conscription would be associated with more violent crime.

The pattern of the correlation between Lottery and Crime_pers is perhaps the most remarkable finding of this enquiry; that départments where the wager is on average high or low have higher crime rates than those départments where the wager is of a medium value, while Creuse sits as an outlier with both a low average wager and low crime rate. Guerry noted this pattern also, but was not able to deduce an explaination: "In public discourse, the lottery is represented as the primary cause of all crimes . . . It is difficult to conceive of the ways in which the lottery, of which we approve as an institution, can give rise to such strong trends."[33] The observed quadratic relationship may be the result of the interplay of other social forces, that may or may not reveal any casual link between the two.

Finally, it is worth remarking on what the variables that are missing from the model reveal. Guerry collected statistics measuring (in his opinion) benevolence, Donations and Donation_clergy. Yet these are not included in the model, and Guerry also did not note any correlation between donations and crime (while he did observe a correlation to education).[34] Again, this is not suprising, and the very notion that such statistics can provide a measure of benevolence I believe is unfounded. Moreover, it would be unreasonable to expect there to be a clear geographic distribution of such traits, though Guerry did not shy away from writing of a France composed of several distinct nations each with their own . . . traditional prejuices."[35] In addition, none of the variables detailing wealth and industry were deemed to be significant predictors of crime against persons, though Wealth is a significant predictor in modelling property crime.[36]

3.2 Limitations of the Model

One of the chief limitations of Guerry's analysis and the data used is that it falls prey to the fallacy of ecological inference. Namely, that relationships that may hold for large groups must also hold for individuals within that group and can always be used for inference about an individual.[37] Guerry's data only allowed him to draw patterns between départments on a national scale, and although he was certainly aware of the limitation of his inferences, he nevertheless confidently pronounced on the differing social and moral characters of the regions of France visable through

ratio infanticides : homicides.

[32] Guerry 1833, pp. 14, 18.

[33] Guerry 1833, pp. 39–40.

[34] Guerry 1833, p. 58.

[35] Guerry 1833; Whitt 2010, pp. 132–134.

[36] Friendly 2007, p. 389.

[37] Firebaugh 2001, p. 4023.

this data.[38]

Similarly, the very ideal of modelling patterns of crime by using social phenomena as predictors needs to be examined. Although the M.S.E. values for the model are not unreasonable (since the mean `Crime_pers` value is 19,754), the low correlation values as detailed in section 1.4 and the relatively low adjusted R^2 value for the model mean that predictons or inferenced based soley on the model should be treated with caution.

Moreover, despite omitting the `Region` variable from the modelling process, the neighbouring départments of Puy-de-Dôme, Creuse and Indre seem to have unusual patterns of crime reporting, based on their outlier status in many of the plots in this enquiry. Given the nature of this data, therefore, a method that incorporated geospatial analysis would be able to model the data with more fidelity.

4 Conclusion

As noted above, this enquiry found significant variables that had escaped Guerry's original analysis, and highlighted the division between the core and periphery through the `Distance` variable. Guerry's hypothesis that crime and education were unrealted was also confirmed, while the insignificance of the vast majority of the variables put doubts on Guerry's statistical project in the first place.

After its publication in 1833, the *Essai* would exert a significant influence on the developing science of sociology and the French government's attempts to impose its rule on the periphery of the country from Paris.[39] Guerry's work, along with that of contemporary statisticians such as Malte-Brun, Dupin and d'Angeville, would give rise to the idea of a division between *France éclairée* ('enlightened France') in the industrialised North, and *France obscure* ('ignorant France') in the agrarian South and West along the St. Malmo-Geneva line, an influential notion for much of the 19th century that was first depicted in Guerry's maps.[40]

[38] Guerry 1833, p. 40; Friendly 2007, p. 378.

[39] Whitt 2010, p. 148.

[40] Guerry 1833, p. 40; Whitt 2010, pp. 132–134; Friendly 2007, p. 372.

References

d'Angeville, A. (1836). *Essai sur la Statistique de la Population française*. Paris: F. Darfour.

Firebaugh, G. (2001). "Ecological Fallacy, Statistics of," in: *International Encyclopedia of the Social and Behavioral Sciences*. Oxford: Pergamon, pp. 4023–4026.

Friendly, M. (2007). "A.-M. Guerry's "Moral Statistics of France": Challenges for Multivariable Spatial Analysis". In: *Statistical Science* 22.3, pp. 368–399.

Friendly, M. and S. Dray (2020). *Maps, Data and Methods Related to Guerry (1833) "Moral Statistics of France"*. Version 1.7.0. URL: https://CRAN.R-project.org/package=Guerry.

Guerry, A.-M. (1833). *Essai sur la statistique morale de la France*. Paris: Crochard. URL: books.google.co.uk/books?id=u3nro2gPONQC.

Hacking, I. (1990). *The Taming of Chance*. Ideas in Context. Cambridge: Cambridge University Press.

James, G. et al. (2013). *An Introduction to Statistical Learning: with Applications in R*. 8th ed. Springer Texts in Statistics. New York: Springer.

Ripley, B. (2020). *boot*. Version 1.3–25. URL: https://CRAN.R-project.org/packages/boot/index.html.

Spiegelhalter, D. (2019). *The Art of Statistics: Learning from Data*. Pelican Books. London: Penguin.

Whitt, H. (2010). "The Civilizing Process and Its Discontents: Suicide and Crimes against Persons in France, 1825–1830". In: *American Journal of Sociology* 116.1, pp. 130–186. URL: https://digitalcommons.unl.edu/sociologyfacpub/496.

Zamoyski, A. (2018). *Napoleon*. William Collins. London: Harper Collins.