BIOSTAT201A Data Analysis Project 2
Brandon Tsai, Lillian Chen, Jackson Chin
All group members contributed equally.
Code will be submitted independently.

Introduction

In the following report, we analyze the County Demographic Information (CDI) dataset. This dataset consists of demographic information collected from 440 counties across the United States, and was collected from 1990-1992; demographics collected for each county include land area, geographical region, population, population age, wealth, and education level of county constituents.

Here, we aim to evaluate several candidate models to determine an appropriate regression model to predict county crime rate from other demographic information available for a county. Using our model, we then evaluate the relationships between these demographic characteristics and county crime rate.

Methods

As we are interested in evaluating predictors for the rate of crime in each county, we first define a new variable, *CRM_1000*. This variable corresponds to the number of crimes per 1000 county population, and is defined by the following equation:

$$CRM\_1000 = crimes \,/\, \left(\frac{pop}{1000}\right)$$

Summary statistics were generated for all informative variables in the CDI data set, with continuous variables summarized by mean and standard deviations, and region summarized by count and percentage (Table 1).

To evaluate the distribution of numeric variables, we plot histograms of each numeric variable in Figure 1. Many of the histograms depicted in Figure 1 demonstrate a strong positive skew. As this skewness may hinder later regression analyses, we evaluate how the log-transformation of these variables influences their distributions. For these log-transformations and all ensuing log-transformations, we use the natural log. Histograms of these log-transformed variables are shown in Figure 2. Python and the *matplotlib* package were used to generate and plot these histograms; *NumPy* and *pandas* were used to log-transform variables and perform data handling, respectively.

To quantify the linear relationships between log-transformed variables, we evaluate the correlation between each pair of log-transformed numeric variables; a heatmap depicting these correlations is shown in Figure 3. These correlations are evaluated using Pearson correlation that assumes we are comparing paired samples from two continuous, independent distributions. Pearson correlation also assumes that the compared samples are derived from normal distributions; we believe this assumption is met from the histograms shown in Figure 2. Python and the *seaborn* package were used to create this heat-map; *NumPy* and *pandas* were used to perform Pearson correlation and handle data.

As a precaution, we also evaluate correlation between log-transformed variables using a Spearman correlation test. Spearman's correlation is non-parametric, meaning that normality in the compared distributions is not required; Spearman correlation is also beneficial in that it can also evaluate non-linear relationships between variables, unlike Pearson correlation that only

evaluates linear relationships between predictors. Python's *SciPy* package was used to derive these Spearman correlation coefficients and resulting p-values.

Further visualization of the data was performed via scatterplots that depict the log-transform of each continuous variable (except for '*hsgrad*') against the natural log of crime rate per 1000 people (Figure 4). As previous analysis projects have shown that demographics can differ significantly across geographic regions, we also provide a scatterplot of the number of beds against crime rate per 1000 people with data points colored by region (Figure 5). These scatterplots were plotted using R and plotting package *BoutrosLab.plotting*.

Simple linear models were performed using the formula *CRM_1000* ~ variable, where all continuous variables, except for '*hsgrad*', were natural log transformed. Stratified linear regression stratified data by geographical region and regressed log(*CRM_1000*) on *beds*. Analysis of crime rate variation across different regions was performed using ANOVA. For the ANOVA, the null hypothesis stated that crime rates were equal across all regions, and the alternative hypothesis stated that crime rate was not equal across all regions. Bonferroni adjusted p-value for pairwise t-tests comparing crime rates per 1000 people by region was calculated using 6 tests (p / n). These ANOVA and simple linear regression models were produced via R.

An initial main effects multiple regression model was created by regressing log-transformed *CRM_1000* on the 13 predictors listed in Table 3a. Pairs of variables with correlation coefficients greater than 0.8 were evaluated for high collinearity and selected variables were dropped to reduce dimensionality and avoid issues of multicollinearity in the model. This final main effects model was then expanded to consider a full model including all 2-way interactions, and this full model was then utilized to perform stepwise regression. Stepwise model selection was performed in both directions using stepAIC from the *MASS* package, and two final models were generated using model selection by the Akaike Information Criterion (AIC) and by the Bayesian Information Criterion (BIC). All tables for multiple regression model summaries were generated using R packages *gt_summary* and *gt*.

P-values less than 0.05 were considered to reach statistical significance, with the exception of the Bonferroni adjusted p-values for pairwise regional comparisons mentioned above.

Results

Prior to performing regression analyses, we evaluate the distributions of our variables to ensure the assumptions imposed by regression are met. Table 1 presents summary statistics of county characteristics. Figure 1 includes histograms for all continuous variables. From these histograms, we find that all of our variables (with exception of *hsgrad*) demonstrate a strong positive skew. This skew could influence our ensuing correlation and regression analyses; Pearson correlation assumes that the compared series are derived from independent normal distributions while regression models require that residual errors are normally distributed.

To form our variables into more-normal distributions, we log-transform each of our numeric variables (except *hsgrad*); the histograms of these log-transformed variables are shown in Figure 2. We find that this log-transformation does shape each variable's distribution to be more normal, suggesting that these log-transformed variables are more appropriate for the following analyses.

To characterize the pre-existing relationships between these variables, we evaluate the Pearson correlation between each pair of numeric variables; again, each of these variables, save *hsgrad*, are log-transformed. The heatmap of correlation coefficients is shown in Figure 3. From this heatmap, we find that the log-transformations of population (ln(*pop*)), the number of non-federal doctors (ln(*docs*)), the number of hospital beds (ln(*beds*)), and total income (ln(*totalinc*)) all demonstrate high correlation with one another. This result is unsurprising; a higher population corresponds to more people that can (1) contribute to the county's total income and (2) increase demand for medical services, therefore leading to larger numbers of doctors and hospital beds.

As for variables that show correlation to crimes per 1000 (ln(*CRM_1000*)), we find that the aforementioned population-correlated variables all demonstrate positive correlations to crimes per 1000 population; none of these individual variables, however, demonstrate a correlation above 0.5, suggesting that none of these variables are effective predictors of crimes per 1000 alone. Of the remaining predictors, we find that poverty rate (ln(*poverty*)) also demonstrates a positive correlation to crimes per 1000, suggesting that poverty rate could be a good predictor for later regression analyses. Additionally, we find that variables like percentage of high school graduates (ln(*hsgrad*)) and bachelor's degree holders (ln(*bagrad*)) demonstrate negative correlation to poverty rate, suggesting they could be interesting predictors by extension.

We also calculated non-parametric ranked Spearman's correlation between the natural log of *CRM_1000* and the natural log of all other continuous variables except 'hsgrad', and the following were statistically significant at alpha = 0.05: pop (Spearman's rho = 0.38, Spearman's rho p-value = 8.6e-17), pop18 (Spearman's rho = 0.26, Spearman's rho p-value = 2.02e-8), docs (Spearman's rho = 0.43, Spearman's rho p-value = 1.06e-21), beds (Spearman's rho = 0.47, Spearman's rho p-value = 1.1e10-25), hsgrad (Spearman's rho = -0.19, Spearman's rho p-value = 4.45e-5), poverty (Spearman's rho = 0.5, Spearman's rho p-value = 2.6e-29), totalinc (Spearman's rho = 0.32, Spearman's rho p-value = 7.98e-12). Although statistically significant, the correlations were modest at best, suggesting that our predictors show significant correlation to the log of CRM_1000 but that no single predictor alone is likely predictive of *CRM_1000*.

ANOVA suggests that the natural log of crime rates per 1000 people are not equal by region ($p < 0.001$) (Table 2). To further explore this, we conducted pairwise t-tests between each region, which resulted in significant differences between all pairwise regions except for regions 3 and 4, using a Bonferroni adjusted p-value cutoff of 0.00833 (Table 3). When looking at scatter plots stratified by region, we can observe the separation of the natural log of *CRM_1000* (Figure 5).

Simple linear regression of the natural log of crime rates per 1000 people on the natural log of continuous variables, except '*hsgrad*', reached statistical significance for *beds*, *poverty*, *docs*, *pop*, *totalinc*, *pop18*, and *hsgrad* (Table 4). When stratifying the linear regression on *beds*, which had the most significant p-value (p=7.79e-25) by region, each independent linear regression was also significant (Table 5).

To search for a best model for county crime rate from the given variables, multiple linear regression was performed on the dataset. The main effects model consisted of the regression of outcome log-transformed *CRM_1000* on 11 predictors after dropping two variables, log-transformed '*pop*' and log-transformed '*docs*', to minimize dimensionality and address

potential issues of collinearity. The main effects model had a multiple R^2 of 0.55 and adjusted R^2 of 0.54 (Table 6). The full model consisting of the main effects and all 2-way interaction terms had a multiple R^2 of 0.68 and adjusted R^2 of 0.60. In the main effects model, covariates *region* and log-transformed *area*, *beds*, *poverty*, and *pcincome* were statistically significant. From the main effects model, the results suggest that a 1% increase in each of *beds*, *poverty*, and *pcincome* is associated with a mean increase of *CRM_1000* by 15%, 36%, and 51%, respectively, and that a 1% increase in *area* is associated with a mean decrease of *CRM_1000* by 5% (Table 6).

Two stepwise regression models were generated from the full model. The final stepwise regression model using AIC had 29 terms and 41 model degrees of freedom, lending itself to a multiple R^2 of 0.65 and adjusted R^2 of 0.62 (Table 7a). The final stepwise regression model using BIC had 16 terms and 18 model degrees of freedom, lending itself to a multiple R^2 of 0.59 and adjusted R^2 of 0.58 (Table 7b). The stepwise model by AIC included all main effects, while the stepwise model by BIC dropped the log-transformed percent population over 65+ years of age. Both models were statistically significant at the α=0.05 level ($F=18.35$, $p<0.001$ for AIC model; $F=34.2$, $p<0.001$ for BIC model). Main effects significant in both stepwise models included log *beds*, *hsgrad*, log *bagrad*, and log *unemp*. 2-way interactions significant in both models included log *beds**log *totalinc*, *hsgrad**log *bagrad*, and log *bagrad**log *pcincome*.

Discussion

Early results in evaluating the distributions of our predictor variables (Figure 1) indicated strong positive skews that we feared could violate the assumptions imposed by downstream regression and correlation analyses. We are able to correct for these skews, however, by log-transforming our variables into more normal distributions (Figure 2). We believe that these transformations improve the normality of our results and improve our later regression model.

Given the modest correlations between the crime rate and other continuous variables, we expect our predictive model to include multiple variables (Figure 3 & 4, Table 3). Even when looking at categorical variables, we observe differences in crime rate by region, as demonstrated by our ANOVA and pairwise t-test of regions (Table 1 & 2). Together, these data suggest that our model will have some complexity in including multiple continuous and categorical variables.

The main effects model had a lower R^2 than both stepwise regression models, which makes sense due to the addition of added terms explaining additional variance. The full model with all main effects and 2-way interactions had a multiple R^2 of 0.68, supporting the expectation from preliminary analyses that the modest pairwise correlations will result in a large model to try to explain variance in the outcome. The two stepwise generated models illustrate the differences in AIC and BIC selection, where BIC imposes a heavier penalty on additional parameters. Additionally, use of AIC usually leads to selection of a model that has high dimensional reality, indicating that there is possible danger of overfitting and that the selected model may not be a true model even though it might have higher correlation. This low penalty on covariate inclusion may explain why certain variables significant in the main effects model are not significant in the AIC stepwise model. On the other hand, while use of BIC may present the danger of underfitting, its use better suits the purpose of finding a true explanatory model since it allows for consistent estimation of the underlying data-generating process. Considering the two

generated models, despite the lower R^2 evident in the BIC model as compared to the AIC model, the BIC model, which has less covariates, may be preferred for the goal of this analysis, which is to find an appropriate model that may explain underlying relationships in the predictors that can better explain variance in the outcome of log(*CRM_1000*).

In reviewing the main effects with significantly nonzero coefficients across both stepwise regression models, log *beds*, log *unemp,* and log *bagrad* demonstrate positive correlations with crime rate per 1000 people while *hsgrad* demonstrates a negative correlation to crime rate per 1000 people. Of these predictors, the correlations of *hsgrad* and log *unemp* with crime rate per 1000 people appear logical from a conceptual perspective; these correlations suggest increasing high school graduation rates and lower unemployment rates correspond to decreased crime rate. The coefficient of log *beds*, however, is somewhat surprising; this predictor corresponds to the natural log of hospital beds in a county and demonstrates a positive correlation to crime rate. We believe, however, that it would be incorrect to conclude that increasing the number of beds in a county would correspond to an increased crime rate per 1000 people as, conceptually speaking, the two variables appear independent. Rather, this correlation may be the result of a confounding variable. In Figure 3, we find that the natural log of hospital beds correlated positively with the natural log of total county income and the natural log of county population, suggesting that the observed correlation between log *beds* and log *CRM_1000* could be the result of some predictor highly correlated to log *beds* instead.

The correlation between the natural log of the percentage of a county's population with bachelor's degrees (log *bagrad*) and crime rate also appears unusual; again, we observe a positive correlation between these two variables (via our stepwise regression analyses), but our prior correlation analyses suggest that a negative correlation should exist instead. In Figure 3, we find that log *bagrad* is negatively correlated with the natural log of unemployment and positively correlated with the percentage of high school graduates; given our understanding of the correlations between high school graduation percentage and unemployment with crime rate per 1000, we would also expect log *bagrad* to be negatively correlated with crime rate per 1000. This positive correlation, therefore, suggests another confounding variable may be involved, and that some external variable correlated with log *bagrad* may be responsible for this positive correlation.

References

Harris, C.R., Millman, K.J., van der Walt, S.J. et al., (2020). Array programming with NumPy. Nature, 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2.

Hunter, J. D., (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90–95.

Iannone, R., Cheng, J., & Schloerke, B., (2021). gt: Easily Create Presentation-Ready Display Tables. R package version 0.3.1. https://CRAN.R-project.org/package=gt

R Core Team, (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Reback et al., (2021). pandas-dev/pandas: Pandas 1.3.4. Zenodo. https://doi.org/10.5281/zenodo.5574486.

Sjoberg, D. D., Curry, M., Hannum, M., Larmarange, J., Whiting, K. & Zabor, E. C. (2021). gtsummary: Presentation-Ready Data Summary and Analytic Result Tables. R package version 1.4.2. https://CRAN.R-project.org/package=gtsummary

Van Rossum, G., & Drake, F. L., (2009). Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.

Venables, W. N. & Ripley, B. D., (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... & Van Mulbregt, P. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods*, *17*(3), 261-272.

Waskom, M. L., (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021, https://doi.org/10.21105/joss.03021

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Figures

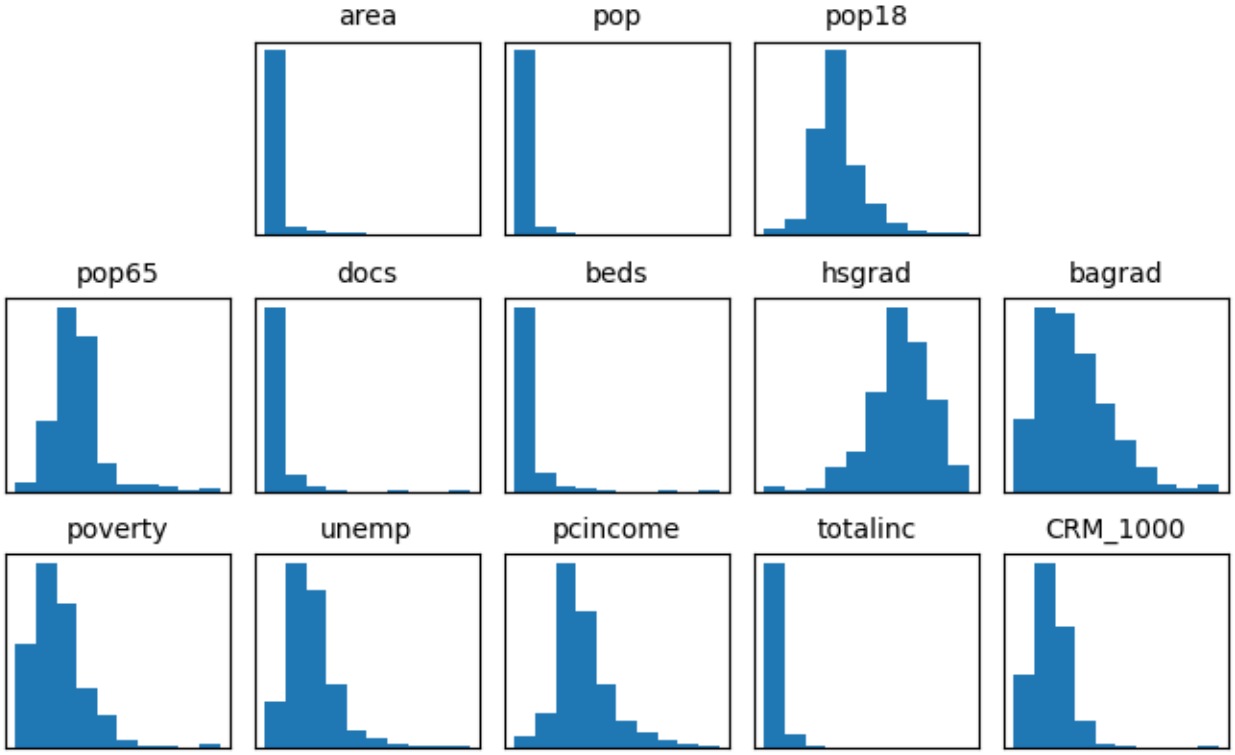

Figure 1: Histograms of each numeric variable.

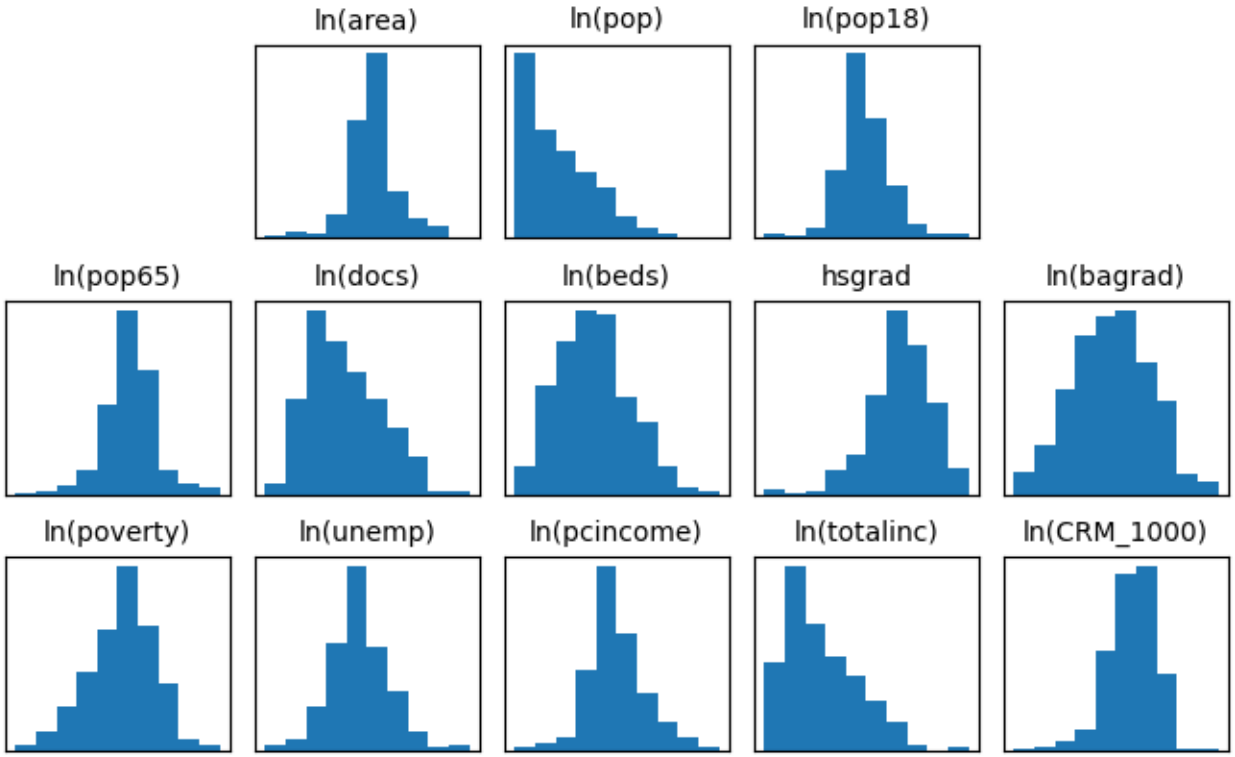

Figure 2: Histograms of each log-transformed numeric variable.

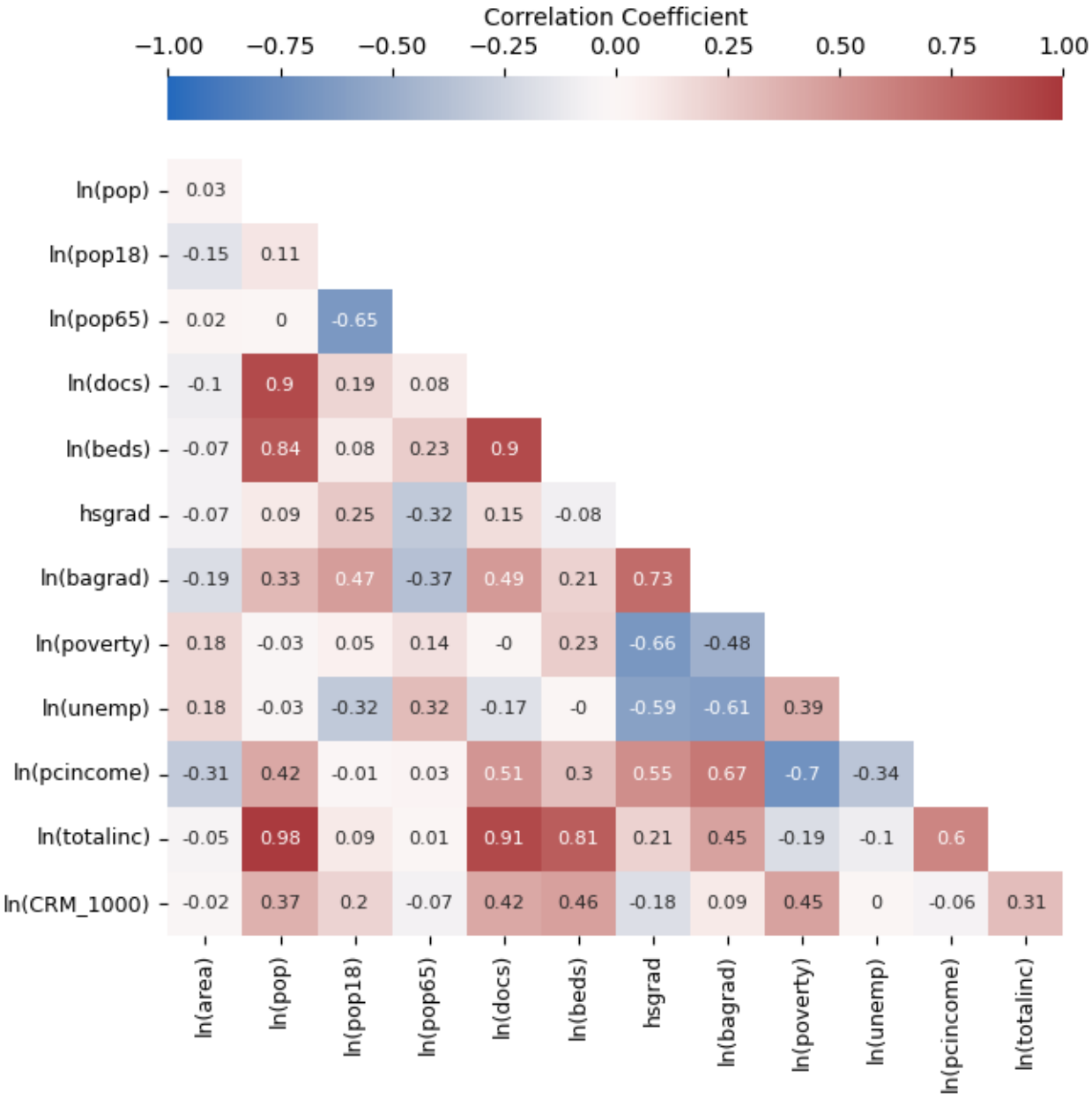

Figure 3: Heatmap of Pearson correlation coefficients between each pair of continuous variables. Each variable, except for 'hsgrad', is log transformed.

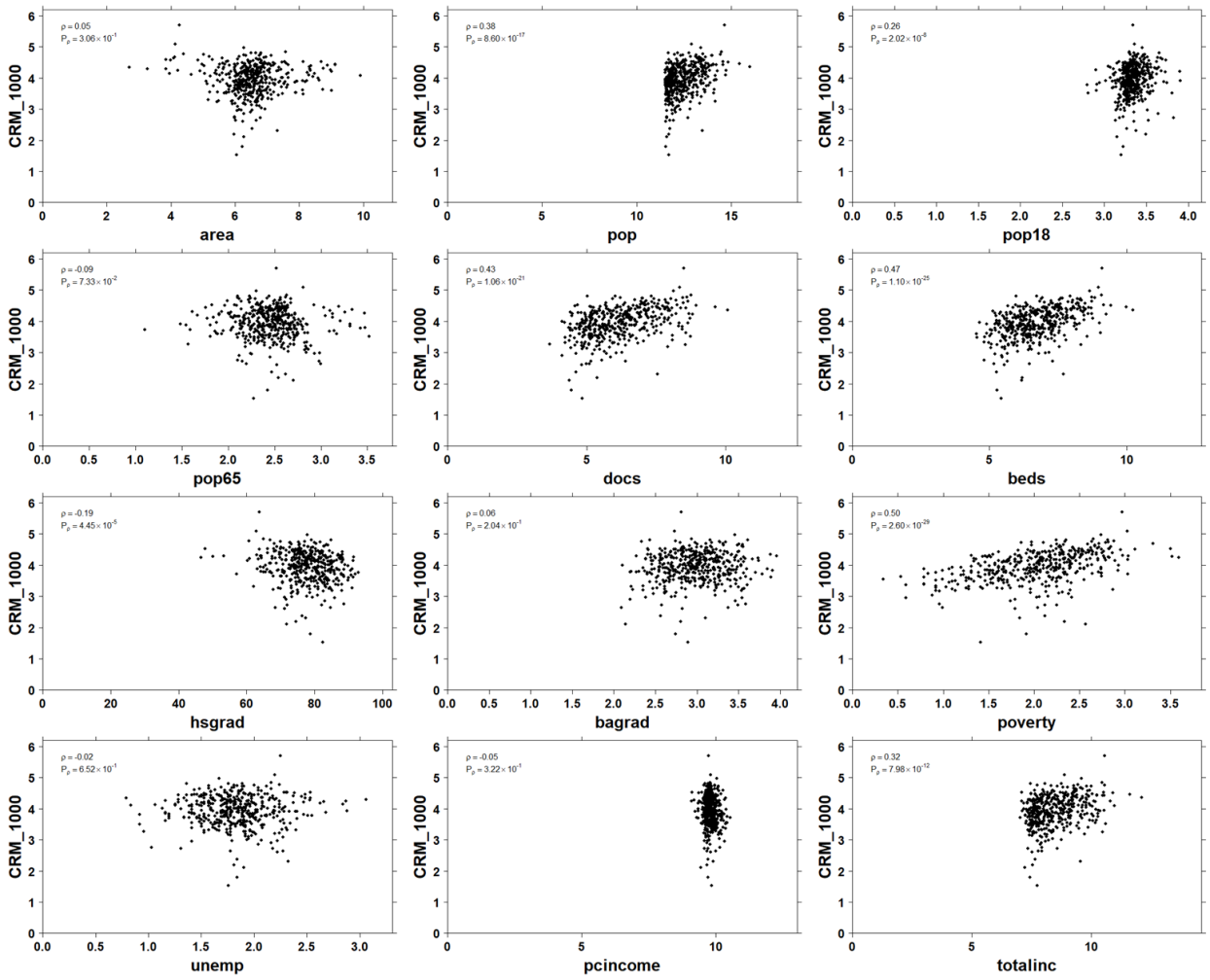

Figure 4. Scatterplots of crime rate per 1000 people by continuous variables. Each variable, except for 'hsgrad', is log transformed.

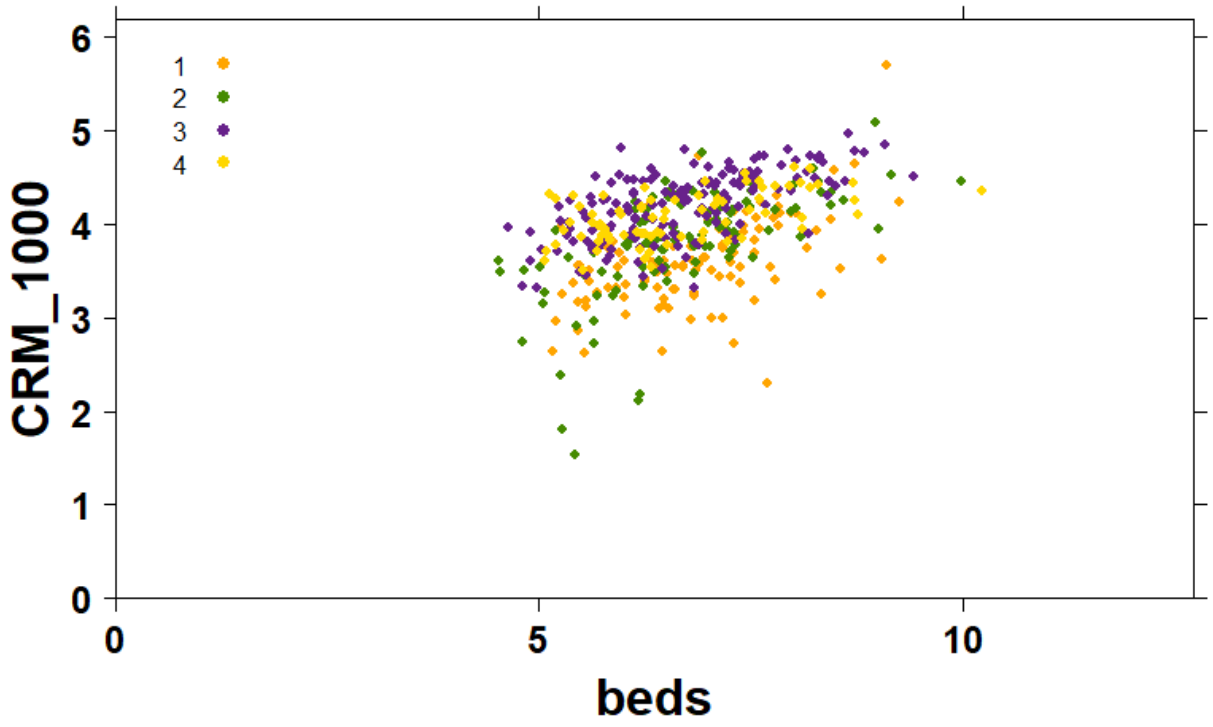

Figure 5. Scatterplot of Scatterplots of the natural log of crime rate per 1000 people by the natural log of beds, stratified by region.

Summary of U.S. County Characteristics

Characteristic	N = 440
Land area (mi), Mean (SD)	1,041 (1,550)
Total population, Mean (SD)	393,011 (601,987)
Percent of population aged 18-34, Mean (SD)	28.6 (4.2)
Percent of population aged 65+, Mean (SD)	12.2 (4.0)
Number of active physicians, Mean (SD)	988 (1,790)
Number of hospital beds, Mean (SD)	1,459 (2,289)
Total serious crimes, Mean (SD)	27,112 (58,238)
Percent high school graduates, Mean (SD)	78 (7)
Percent bachelor's degrees, Mean (SD)	21 (8)
Percent below poverty level, Mean (SD)	8.7 (4.7)
Percent unemployment, Mean (SD)	6.60 (2.34)
Per capita income (USD), Mean (SD)	18,561 (4,059)
Total personal income (millions USD), Mean (SD)	7,869 (12,884)
Geographic region, n (%)	
Northeast	103 (23)
North Central	108 (25)
South	152 (35)
West	77 (18)

Table 1. Summary of U.S. County Characteristics from the CDI data set.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
region	3	26.7	8.9	45.94	<2e-16
Residuals	436	84.47	0.194		

Table 2. Summary table of ANOVA of the natural log of crime rates per 1000 people by region.

	1	2	3
2	0.00028	-	-
3	< 2e-16	4.90E-12	-
4	5.00E-13	4.10E-05	0.04901

Table 3. P-values of pairwise t-tests comparing the natural log of crime rates per 1000 people by region. Regions 1-4 correspond to Northeast, North Central, South, and West regions of the U.S., respectively.

Simple Linear Regression of Crime Rate per 1000 Population on Each Variable		
Variable	Coefficient	pvalue
Log land area (mi)	-0.014	0.621
Log total population	0.235	<0.001
Log percent of population aged 18-34	0.723	<0.001
Log percent of population 65+	-0.115	0.149
Log number of active physicians	0.186	<0.001
Log number of hospital beds	0.233	<0.001
Percent high school graduates	-0.013	<0.001
Log percent bachelor's degrees	0.128	0.059
Log percent below poverty level	0.427	<0.001
Log percent unemployment	0.004	0.956
Log per capita income (USD)	-0.143	0.218
Log total personal income (millions USD)	0.174	<0.001
North Central (ref: Northeast)	0.222	<0.001
South (ref: Northeast)	0.616	<0.001
West (ref: Northeast)	0.494	<0.001

Table 4. Summary table of linear modeling of log crime rate per 1000 people by the log of continuous variables except for percent high school graduates and region, crime rate ~ variable.

Region	p-value
1	8.65E-09
2	1.77E-12
3	<2e-16
4	5.82E-07

Table 5. P-values from linear regression of the natural log of crime rates per 1000 people on the natural log of beds, stratified by region. Regions 1-4 correspond to Northeast, North Central, South, and West regions of the U.S., respectively.

Main Effects Multiple Regression Model of Log Crime Rate on 11 Predictors

Characteristic	Beta (95% CI)[1]	p-value
Log land area (mi)	-0.05 (-0.10 to -0.01)	**0.029**
Log percent of population aged 18-34	0.36 (-0.01 to 0.74)	0.058
Log percent of population aged 65+	-0.07 (-0.25 to 0.10)	0.42
Log number of hospital beds	0.15 (0.06 to 0.24)	**<0.001**
Percent high school graduates	0.00 (-0.01 to 0.01)	0.86
Log percent bachelor's degrees	-0.01 (-0.24 to 0.21)	0.91
Log percent below poverty level	0.36 (0.22 to 0.51)	**<0.001**
Log percent unemployment	0.08 (-0.06 to 0.23)	0.25
Log per capita income (USD)	0.51 (0.12 to 0.90)	**0.011**
Log total personal income (millions USD)	0.02 (-0.08 to 0.13)	0.64
Geographic region		
Northeast	—	
North Central	0.22 (0.11 to 0.33)	**<0.001**
South	0.55 (0.45 to 0.66)	**<0.001**
West	0.50 (0.37 to 0.63)	**<0.001**

[1] CI = Confidence Interval

Table 6. Main effects multiple regression model of log crime rate per 1000 people on 11 predictors. Predictors log total population and log number of active physicians were excluded to reduce dimensionality and multicollinearity.

Characteristic	Beta (95% CI)[1]	p-value
Log land area (mi)	1.5 (-0.57 to 3.6)	0.15
Log percent of population aged 18-34	-14 (-28 to 0.02)	0.050
Log percent of population aged 65+	2.6 (0.18 to 5.0)	**0.035**
Log number of hospital beds	0.69 (0.33 to 1.1)	**<0.001**
Percent high school graduates	-0.34 (-0.65 to -0.02)	**0.035**
Log percent bachelor's degrees	21 (12 to 29)	**<0.001**
Log percent below poverty level	-0.14 (-0.87 to 0.59)	0.71
Log percent unemployment	5.5 (1.8 to 9.3)	**0.004**
Log per capita income (USD)	-3.2 (-9.2 to 2.9)	0.30
Log total personal income (millions USD)	0.15 (-0.16 to 0.47)	0.34
Geographic region		
Northeast	—	
North Central	-4.5 (-9.7 to 0.60)	0.083
South	-4.4 (-8.9 to 0.17)	0.059
West	2.9 (-2.6 to 8.4)	0.30
Log land area (mi) * Log percent of population aged 18-34	0.37 (0.04 to 0.71)	**0.030**
Log land area (mi) * Log percent unemployment	-0.18 (-0.32 to -0.03)	**0.018**
Log land area (mi) * Log per capita income (USD)	-0.26 (-0.45 to -0.06)	**0.010**
Log percent of population aged 18-34 * Log percent of population aged 65+	-1.0 (-1.7 to -0.23)	**0.009**
Log percent of population aged 18-34 * Log percent bachelor's degrees	-2.4 (-3.5 to -1.2)	**<0.001**
Log percent of population aged 18-34 * Log percent unemployment	-1.4 (-2.5 to -0.37)	**0.008**
Log percent of population aged 18-34 * Log per capita income (USD)	2.5 (0.85 to 4.2)	**0.003**
Log percent of population aged 18-34 * Geographic region		
Log percent of population aged 18-34 * North Central	-0.50 (-1.5 to 0.51)	0.33
Log percent of population aged 18-34 * South	0.24 (-0.74 to 1.2)	0.63
Log percent of population aged 18-34 * West	-1.2 (-2.4 to 0.01)	0.052
Log percent of population aged 65+ * Geographic region		
Log percent of population aged 65+ * North Central	0.62 (-0.01 to 1.3)	0.052
Log percent of population aged 65+ * South	0.77 (0.25 to 1.3)	**0.004**
Log percent of population aged 65+ * West	0.50 (-0.13 to 1.1)	0.12
Log number of hospital beds * Log percent below poverty level	-0.08 (-0.18 to 0.01)	0.095
Log number of hospital beds * Log total personal income (millions USD)	-0.05 (-0.08 to -0.02)	**0.004**
Log number of hospital beds * Geographic region		
Log number of hospital beds * North Central	0.07 (-0.03 to 0.17)	0.17
Log number of hospital beds * South	-0.07 (-0.17 to 0.02)	0.14
Log number of hospital beds * West	-0.06 (-0.18 to 0.05)	0.25
Percent high school graduates * Log percent bachelor's degrees	0.03 (0.01 to 0.05)	**0.010**
Percent high school graduates * Geographic region		
Percent high school graduates * North Central	0.04 (0.01 to 0.06)	**0.001**
Percent high school graduates * South	0.02 (0.00 to 0.04)	**0.014**
Percent high school graduates * West	0.00 (-0.02 to 0.02)	0.88
Log percent bachelor's degrees * Log per capita income (USD)	-1.5 (-2.3 to -0.76)	**<0.001**
Log percent below poverty level * Log total personal income (millions USD)	0.14 (0.02 to 0.27)	**0.028**

Log percent unemployment * Geographic region		
Log percent unemployment * North Central	0.72 (0.32 to 1.1)	**<0.001**
Log percent unemployment * South	0.52 (0.12 to 0.91)	**0.010**
Log percent unemployment * West	0.28 (-0.18 to 0.74)	0.23
Percent high school graduates * Log per capita income (USD)	0.02 (-0.01 to 0.06)	0.16

[1] CI = Confidence Interval

This model has a multiple R-squared of 0.65 and adjusted R-squared of 0.62

F-statistic: 18.35 on 41 and 398 DF, p-value: < 0.001

Table 7a. Stepwise regression model including main effects and 2-way interactions with variable selection by AIC.

Characteristic	**Beta (95% CI)[1]**	**p-value**
Log land area (mi)	0.33 (0.03 to 0.64)	**0.033**
Log percent of population aged 18-34	0.60 (0.29 to 0.91)	**<0.001**
Log number of hospital beds	0.84 (0.50 to 1.2)	**<0.001**
Percent high school graduates	-0.06 (-0.11 to -0.01)	**0.012**
Log percent bachelor's degrees	9.2 (4.6 to 14)	**<0.001**
Log percent below poverty level	-0.64 (-1.3 to 0.02)	0.057
Log percent unemployment	1.3 (0.44 to 2.2)	**0.003**
Log per capita income (USD)	4.2 (2.3 to 6.0)	**<0.001**
Log total personal income (millions USD)	0.38 (0.11 to 0.64)	**0.006**
Geographic region		
Northeast	—	
North Central	0.24 (0.14 to 0.34)	**<0.001**
South	0.61 (0.51 to 0.71)	**<0.001**
West	0.52 (0.39 to 0.65)	**<0.001**
Log land area (mi) * Log number of hospital beds	-0.06 (-0.09 to -0.02)	**0.003**
Log land area (mi) * Log percent below poverty level	0.16 (0.06 to 0.26)	**0.002**
Log land area (mi) * Log percent unemployment	-0.18 (-0.31 to -0.06)	**0.005**
Log number of hospital beds * Log total personal income (millions USD)	-0.04 (-0.07 to -0.01)	**0.012**
Percent high school graduates * Log percent bachelor's degrees	0.02 (0.01 to 0.04)	**0.011**
Log percent bachelor's degrees * Log per capita income (USD)	-1.1 (-1.7 to -0.58)	**<0.001**

[1] CI = Confidence Interval

This model has a multiple R-squared of 0.59 and adjusted R-squared of 0.58

F-statistic: 34.2 on 18 and 421 DF, p-value: < 0.001

Table 7b. Stepwise regression model including main effects and 2-way interactions with variable selection by BIC.