Analysis of Continuous Data project

Thomas Sertijn, Ilja Van Bever, Lieselot Van de Putte

2025-11-09

Research question

For this research the effect of socio-economic disadvantage to violent crime rates was investigated. More specifically we want to explore the association between poverty and violent crime rates in the USA. Becker (1968) stated that the decision to commit crime is a rational choice where people weigh the benefits and costs of committing crime against each other. It could be argued that the incentive to commit crime is higher for people who have a lower income, as they have more to gain and less to lose by committing crimes. Following this, we would then also expect that communities with higher poverty rates will also be associated with higher crime rates. Depending on the results of our analysis, our analysis could be used to inform relevant policies. It would, for instance, give another argument to redistributive policies: if an effect is found, this policy might also result in a reduction in violent crime, next to an economic benefit. Our analysis hopes to shed further light on this issue.

Methods

Design of the dataset

For this research, a dataset related to data gathered for the prediction of serious crime rates in the US. This dataset combines 1990 U.S. Census socio-economic data, 1990 law enforcement data from the Law Enforcement Management and Admin Stats (LEMAS) survey, and 1995 FBI crime data, thereby creating two cohorts. For the FBI crime data, it is mentioned that states with a lower amount of visitors have a lower per capita crime rate and vice verse. The LEMAS survey covers all communities with police departments of at least 100 officers and a random sample of smaller departments. If communities were absent from either the crime or census datasets (e.g., those with very small departments), then they were removed. All demographic data is from 1990, but per-capita crime rates use 1995 population counts. Finally, rape counts, a component of violent crime, are missing in some states due to inconsistent reporting, which resulted in missing total violent crime values for those states. We will investigate whether this missingness has probably a large effect on the model and if necessary use imputations.

our work

For the purpose of our research question we selected a subset of variables present in in the dataset. Some economic variables: PctPopUnderPov, perCapInc, and PctEmploy, which give some information about economic status of the community, and a sociological subset, containing education levels in the communities, combined with NummImmig, RacePctBlack, AgePct12t29, all variables giving some info about the social composition and demographic structure of the communities. These variables were then used to anwser the following questions:

1. What is the association between poverty rates (*PctPopUnderPov*) and violent crime rates (*ViolentCrimesPerPop*) in U.S. communities? We first present an univariate analysis to gain an initial understanding of whether our hypothesis holds, how strong the relationship is, but do not take any confounding factors into account yet.

2. How does the found association change when we include other socio-economic factors in our analysis? To assess this, we use a multivariate regression to analyze if the relationship is confounded by other socio-economic variables (1), whether there are relevant interaction effects at play (2) and to see if the inclusion of these other socio-economic variables improves the performance of our model substantially (3).

model building

Before building the models, some descriptive analysis of the dataset and variables was performed. Next, the dataset is randomly split into a training set (80% of the data) and a holdout set (20% of the data). This holdout set will be used to validate the final model. For each model, the model is trained on the train set and assumptions of linear regression are checked. For the multivariate model, an All-variable procedure was emplored to test every combination of possible variables as a model. The best model will be chosen based on the Bayesian Information Criterion. After which the assumptions of this best model are checked. Next, partial residual plots are generated to check if each of the added variables are in the correct functional form and whether transformations are needed to better capture their relationship with the outcome. Afterwards, the interaction term with the main predictor PctPopUnderPov is selected based on this same BIC procedure. As a final part of the model building, the multicollinearity is assess using the Variance Inflation Factor (VIF).

After determining the final model, some model diagnostics are calculated to determine outliers and influential points. After diagnostics, the model is validated by the holdout dataset.

Data preparation

After loading in the data and selecting the subset, *NumImmig* is converted to a percentage by dividing it by the population size and multiplying by 100% since the outcome variable *ViolentCrimesPerPop* (total number of violent crimes per 100K population) is expressed relative to the population size. We call this converted variable *PctImmig*. It's important to mention that this is not an exact transformation, because all demographic data is from 1990, but per-capita crime rates use 1995 population counts.

By examining the missing data, there were 221 (of the 1994) observations identified as NA values for the outcome variable *ViolentCrimesPerPop*. Imputations in this case would not be very helpful as the model would then partially be trained on artificial outcomes. As already mentioned in the section on the study design, these NA values are possibly due to the fact that rape counts, a component of violent crime, were not included in the statistics for several states. The rows where the outcome variable has an NA value are removed, as these rows are not useful for the regression. It can be noted that the variables *countyCode* and *communityCode* are also frequently unknown.

Univariate descriptives

After removing NA values from the database univariate descriptives are calculated, for both the missing values and the non-missing values.

To gain insight into the univariate distributions, boxplots and histograms are generated. These histograms are then compared to the histograms of the distributions for the variables with missing data for violent crimes to evaluate if there are important deviations. Neither the summary statistics nor the histograms indicate that the missing data have characteristics that differ substantially from the non-missing data.

Multivariate descriptives

After investigating the univariate descriptives, multivariate descriptives are looked at. The constructed correlation matrix shows the extent to which the variables in the dataset are correlated with each other. It's important to mention that these correlations are indicators of an association, not of a causation. The

following predictors are highly correlated with each other. Therefore, it might be best not to include them together in a model later.

- *PctNotHSGrad* and *PctLess9thGrade* ($r = 0.93$)
- *perCapInc* and *PctBSorMore* ($r = 0.77$)
- *PctNotHSGrad* and *PctBSorMore* ($r = -0.75$)

However the choice for predictors for the model will be dealt with thoroughly during the model building.

It is noticeable that the variable *racepctblack* is the one most strongly correlated with the outcome variable ($r = 0.63$), even more than *PctPopUnderPov*, the head predictor that was chosen for this research.

The following scatter plots were generated:

- for each variable, a scatter plot showing the relationship with the outcome variable *ViolentCrimesPerPop*;
- for each variable, a scatter plot showing the relationship with the main predictor variable *PctPopUnderPov*.

The scatter plots showed that most variables display a roughly linear relationship with *ViolentCrimesPerPop*, although that trend is often distorted in the extreme regions of the x-asis. It's also visible that some variables exhibit a linear relationship with *PctPopUnderPov*. This suggests that these variables are probably not suitable as additional predictors when *PctPopUnderPov* is already included in the model as this may introduce multicollinearity.

The scatter plots also reveal one outlier for the *ViolentCrimesPerPop* variable. However, outliers and their influence will be further investigated in the diagnostic section.

It was also investigated whether or not small communities have a higher probability to have more extreme values of the response and predictor variables, scatter plots were created with log(*population*) as x variable (see appendix). The resulting scatter plots show that communities with a very small population indeed show a very large spread for all variables.

Model Building

Univariate linear regression

The simple univariate regression equation we estimate with the training set is given as follows:

$$ViolentCrimesPerPop_i = \beta_0 + \beta_1 \cdot PctPopUnderPov_i + \epsilon_i$$

Here a coefficient of 38.87 could be found which represents the expected increase in violent crimes per 100K population if poverty rate increases by one percentage point. The R-squared value of ‘r round(summary(fit_simple)$r.squared, 4) represents the proportionate reduction of total variation in *ViolentCrimesPerPop* by the poverty percentage, in this univariate model this R-squared is rather small and can be taken to mean that poverty by itself is insufficient to explain the variation in violent crime rates. For the simple linear model the assumptions of linearity are violated. Larger outcome values tend to be underestimated, the variance for larger outcome values is larger and the QQ-plot shows violation of the normality assumption. It could also be seen that larger populations tended to have larger residuals meaning that the model tends to underestimate the total number of crimes for large populations.

In the next step, we extend the model by adding relevant predictors and reassess the assumptions.

Multivariate model

The first all-variable procedure resulted as model in:

$$ViolentCrimesPerPop_i = \beta_0 + \beta_1 \cdot PctPopUnderPov_i + \beta_2 \cdot PctLess9thGrade_i + \beta_3 \cdot PctNotHSGrad_i + \beta_4 \cdot PctImmig_i + \epsilon_i$$

However for this model, the assumptions of normality and equal error variances of the error terms are again violated.

Applying the log transformation offered a substantial improvement of our model showing approximate constant variances and residuals reasonably close to normal. The reduction in R-squared suggests that the model-fit was inflated due to heteroscedasticity.

Using the log transform Y variable to selected a new model resulted in a similar model as before but with the PctBSorMore variable instead of the PctNotHSGrad. In this model, it could be observed from the partial regression plots that PctRaceBlack was not in the correct functional form. For this reason, a logit transformation was attempted which resulted in a more linear relation on the partial regression plots. Fitting the model with this logit transformed racepctblack seemed to also stabilize the residual vs fitted plot. This was followed by selecting a interaction term for the model, which showed that the best interaction given the BIC values would be between the main predictor PctPopUnderPov and PctLess9thGrade. However, this interaction introduced a high Vif value for both predictors and the interaction term possibly due to the presence of multiple education variables. Opting us to only take one education variable going forward. This resulted in a similar model with the PctLess9thGrade changed for agePct12t29. This model showed to have as best interaction term PctPopUnderPov:agePct12t29. However, This interaction resulted in a High VIF fot both variables and their interaction term. For that reason, the second best interaction term PctPopUnderPov:PctBSorMore was chosen in the final model. This resulted in a final model:

$$ViolentCrimesPerPop_i = \beta_0 + \beta_1 \cdot PctPopUnderPov_i + \beta_2 \cdot PctBSorMore + \beta_3 \cdot agePct12t29 + \beta_4 \cdot PctImmig_i + \beta_5 \cdot race_b lack_l ogi$$

Model diagnostics

After deciding on the model and the form of the variables, the effect of the outliers was determined. This was done based on the Deleted Studentized residuals, the Coook's distance, the leverage, DFFits, and DFBetas. These values were plotted, and values crossing a certain threshold were visualized and seen as influential. If certain datapoints were seen in multiple plots, then these were most likely more influential and possibly problematic outliers.

Using these plots, many influential outliers could be discovered. In order to exclude any datapoints that were influential due to observation errors, which should be removed form the dataset. The variable values of these points were extracted and manually checked, to check whether these are observations errors, rare cases, or valid but extreme cases.

By looking at all 318 Communities in the dataset that were flagged as problematic, so crossing the treshold of one of the diagnotic tests, no evidence could be found of no evidence could be found of errors or anomalies that would necessitate removal. Most of these communities were small and exibited extreme demographic or crime-related characteristics. For example, two communities, Martinsvillecity and Vidorcity, that were flagged by all 5 diagnostic tests, were reporeted to not have any people of black color in there community. As these are both small communities, we have no reason to suspect this observation to be wrongly annotated. Another example is Spencercity, which was reported to have a ViolentCrimesPerPop of 0, while this is not common in our dataset, we lack evidence to say that this we lack evidence to say that this observation is erroneous. Other observations showed either a high or low value in one or more variables compared to the majority of communities, reflecting genuine heterogeneity rather than data quality issues.

These findings show that these flagged communities likely represent valid but extreme cases present in our dataset whose demographic or crime-related profiles differ substantially from the average. Therefore, there is no reason to discard any of them. So instead of removing them, another way to account for their impact would be to run a robust regression, instead of a normal linear regression.

Attemping to solve the influencial outliers with robust regression.

To futher assess and account for the impact of these points, we fit a robust regression using Bisquare weights.

When using a robust model, the coefficients revealed nearly identical to those from the OLS, indicating that the influential points are not the result of gross errors. The small differnces in coefficients show that these are already quite stable and thus that maybe the outliers and influential points did not distort the OLS model significantly

interpretation final model

Our final model is:~, logViolent, PctPopUnderPov + PctBSorMore + race_black_logit + agePct12t29 + PctImmig + PctPopUnderPov:PctBSorMore.

Notes on interpretation The intercept on its own is not interpretable as it would require all predictor variables (and interaction terms) to be zero. This is not in the scope of our model. If it were, the intercept of NA could be interpreted as the estimated log of expected (violent crime rate per 100k + 1) when all predictor variables are 0. There is no use in interpreting interaction-term coefficients on their own, as they indicate the effect of one predictor on the other, but have no direct effect on its own.

Interpretation of PctPopUnderPov The effect of PctPopUnderPov varies with PctBsorMore through their interaction. This means that the effect of PctPopUnderPov is not simply given by its coefficient, but instead we can look at the estimated effect of PctPopUnderPov on the mean of the logarithm of (violent crime rate per 100k + 1) as $0.0258367 + 8.0268453 \times 10^{-4}$ * PctBSorMore. If we take for example the mean of the PctBSorMore to calculate the effect of PctPoponderpov when the interacting variable is held at its average and all variables are held constant, we can find 0.0442439. The confidence interval of this effect is [0.0377381, 0.0515864] With a confidence coefficient of .95 we estimate that the true conditional effect of PctPopUnderPov (with its interaction) on the expected value of the logarithm of (violent crimes per population of 100k + 1) per unit increase in PctPopUnderPov (1 percent point) is somewhere between 0.0377381% and 0.0515864%, when the interacting variable is held at its average and all other variables are held constant. The interaction effect is reinforcing: with increasing percentages of people with higher education, the effect of PctPopUnderPov on the logarithm of (violent crimes + 1) gets larger. Theoretically this could be explained due to the bigger gap in socio-economic standing, which could provoke more violent crimes.

Interpretation of PctBSorMore The interpretation for PctBSorMore is similar to the one above because of their interaction. The conditional effect on of PctBSorMore on the log violent crime rate is given by: $-0.0259593 + 8.0268453 \times 10^{-4}$* PctPopUnderPov. Similarly to before we can take the mean of PctPopUnderPov and estimate the conditional effect of PctBSorMore on the log expected value of crime rate per unit increase of PctBSorMore, when keeping poverty at its mean and all other variables constant: -0.0259593 + 0.0092194. Contrary to above, we now see a interference effect: for higher levels of PctPopUnderPov the negative effect of PctBSorMore becomes less negative (the slope becomes less steep). With a confidence interval of [-0.0206272, -0.0124176] which can be interpreted as: with a confidence coefficient of 0.95 we estimated that the true effect on the logarithm of (expected crime rate + 1) by increasing PctBSorMore by one unit (1%), keeping PctPovUnderPop at its mean and all other variables constant is somewhere between [-0.0206272, -0.0124176]. Thus, under the conditions stated above, if a higher percentage of the population has a Bachelors degree or higher, the estimated logarithm of the (mean crime rate + 1) decreases and this effect is less pronounced for higher levels of PctPopUnderPov.

Interpretation of race_black_logit The interpretation of race_black_logit is more complicated as on top of the log transformed response variable, we now also have a logit transformed predictor variable. It has no interaction terms so we can interpret the main effect on its own. A one-unit increase in the log-odds of percentage African American is associated with a 0.2734775 change in the logarithm of the (expected value of Violent Crimes per 100k + 1) when keeping all other variables constant; where a one-unit increase in the log-odds has to be interpreted as a multiplication of the odds of being African American by e ($\approx$ 2.72). 95% confidence interval is [0.2506591, 0.2962959] respectively. With a confidence coefficient of 0.95 we estimated that the true parameter of race_black_logit is somewhere between 0.2506591 and 0.2962959. This

is a positive association: for increasing levels of the log-odds of percentage African Americans, there is an increase in the logarithm of the expected value of (violent crime rates per 100k + 1).

Interpretation of PctImmig – nog niet af The effect of PctImmig on log Violent crime rates is: 0.0300765. With confidence coefficient .95 we estimate that the percentage change in expected crime rates per 100k population per unit increase in PctImmig when keeping all other variables constant will be somewhere between 0.0256521 and 0.0345009. This increase in violent crimes associated with the increase in percentage of foreign borns agrees with the ....<????>

agePct12t29 – nog niet af For the percentage of people between the age of 12 and 29, we find a negative association with the logartihm of (violent crime rates per 100k + 1). This is expressed by a coefficient of -0.0256855 with 95% confidence interval [-0.0348363, -0.0165347]. with 95% confidence we estimate that the true parameter of agePct12t29 is somewhere between -0.0348363 and -0.0165347. This negative association is not in line with our prior assumption.

Family confidence interval

Summary

Dusja multivariate model stuk beter dan univariate model als je kijkt naar de tabel

model validation

refitting the model on the test data

prediction of the test data

Using the holdout set, we compute the Mean Squared Prediction Error (MSPR) and comparing it to the Mean Squared Error (MSPR) when the model is used to predict data in the training set.

As expected, 95% of the data points fall within the 95% prediction intervals. However, after backtransformation of the outcome variables it can be noticed that the prediction intervals are far too wide, making the predictions difficult to use.

Resultaten zijn niet zo goed, dus dit nog niet herwerkt.

Statistical discussion

There are several characteristics of the dataset that require discussion. First, we focus on how the missing data in our dataset affects our found results. The main source of missing data in our analysis, is the ViolentCrimesPerPop variable, where 221 observations are missing because of incomplete rape reporting in Midwestern States. This missingness could lead to biased estimates if these communities differ in crime-related or demographic characteristics. Therefore, we tested to check if the characteristics of the communities with missing data were different from those of the non-missing data, and found no important differences. We thus assume that this missingness does not have an important effect on our found results. However, if these communities are different in ways not captured by our measured variables, there could still be bias present.

Second, the data from the LEMAS survey consists of all communities with police departments with more than 100 officers, but only a random sample of smaller ones. Because of this, large (and thus more urban) communities are overrepresented, while small communities are underrepresented in the dataset. Thus, the relationships we find in the analysis are more likely to hold in the context of larger cities, and may be less likely to translate to rural communities.

Additionally, based on our dataset, we expected that smaller communities have greater variance in both predictor and outcome variables, which we see confirmed in our scatter plots. The reason for this is that per capita crime rates in small populations are more volatile, as one crime in a community of 1,000 translates to

100 per 100,000 population. Because of this, small changes in actual crime numbers lead to large fluctuations in converted rates. We dealt with this heteroskedasticity by applying a log transformation. Nevertheless, variability for these smaller communities should thus be interpreted with caution.

Another caveat is that the FBI's data is from 1995, while the Census' populations data containing demographic variables are from 1990. This way, data from 1990 is used to predict crime rates from 1995. If the demographic variables of communities changed drastically during this period, the predictor variables would not represent these communities well in 1995, when the actual crimes happened. For instance, if we assume that the poverty rate would increase substantially between 1990 and 1995, the 1990 poverty rate used in our analysis might then underestimate the actual relationship. However, we expect that, because of the short timespan and as we expect most of the socio-economic variables we used to change only slowly, this to have a limited impact on our results.

Then, it should be noted that the data are from 1990-1995. Since then, the social and economic environment have changed, implying that our estimated relationships would not hold today as crime rates, reporting practices,... have evolved. Thus, these results should be carefully intepreted in today's context as the mechanisms driving these relationships are different in comparison to when the data collection happened.

At last, it is mentioned in the description of the original dataset that many relevant factors are not included. Specifically, it is mentioned that per capita crime rates are calculated using resident populations, so communities with large numbers of visitors are expected to have higher crime rates, even if the actual risk is not different across communities. This, together with other unmeasured factors, implies that our results should be interpreted as correlations and not as causations.

We conclude that our regression analysis results are thus valid for a subset of U.S. communities in the 1990's and that the results should be cautiously interpreted in this context.

Final discussion/conclusion LINEAR MODEL

In this report, we investigate how violent crime rates and socio-economic disadvantage, and more specifically the poverty rate, are associated in U.S. communities. We do so by using a merged dataset with detailed data on both crimes and demographic variables from 1990 and 1995. Specifically, our analysis consists of two parts. First, we perform a univariate linear regression to investigate the direct relationship between poverty and violent crime rates. Here, we find a significant positive relationship. However, we want to assess whether other socio-economic variables on top of poverty can substantially improve our model (1), whether these predictors can account for confounding (2), and whether there are interaction effects at play (3).

Therefore, we make use of a multivariate model including a range of other (socio-economic) predictor variables. We find that this final model indeed does substantially improve our model in comparison to the univariate model. Notably, we also include an interaction term in this final model between percentage of residents with bachelor's degrees or higher and the poverty rate. The positive coefficient on this interaction term indicates that the impact of poverty on crime rates depends on the percentage of residents holding a bachelor's degree or more, implying that the effect of poverty on violent crime becomes more positive as education increases. We assume that this is because a high poverty rate in combination with a high educational attainment rate could imply there to be the a large level of inequality present in the community. This inequality then may be associated with higher violent crime rates. There is a huge literature on this topic, which would be impossible to completely summarize here. However, the interested reader is referred for this purpose to Kelly (2000), Fajnzylber et al. (2002) and Kang (2016), confirming this to be a plausible explanation.

Several limitations should be highlighted though. The dataset we use overrepresents larger communities, contains missing data and uses outdated demographic predictors. It is also doubtful whether the results would still hold in today's context because of demographic changes. Interpreting the results in this context, our results imply that policies aiming to reduce poverty could reduce violent crime rates, especially in combination with policies that reduce (educational) inequalities. However, it remains important that these results should be interpreted as correlational and not as causal, as many confounding factors that our analysis cannot for, could still be at play.

References

Becker GS (1968) Crime and Punishment: An Economic Approach. Journal of Political Economy 76: 169–217

Fajnzylber, P., Lederman, D., & Loayza, N. (2002). Inequality and violent crime. The journal of Law and Economics, 45(1), 1-39.

Kang, S. (2016). Inequality and crime revisited: effects of local inequality and economic segregation on crime. Journal of Population Economics, 29(2), 593-626.

Kelly, M. (2000). Inequality and crime. Review of economics and Statistics, 82(4), 530-539.

References dataset

U. S. Department of Commerce, Bureau of the Census, Census Of Population And Housing 1990 United States: Summary Tape File 1a & 3a (Computer Files),

U.S. Department Of Commerce, Bureau Of The Census Producer, Washington, DC and Inter-university Consortium for Political and Social Research Ann Arbor, Michigan. (1992)

U.S. Department of Justice, Bureau of Justice Statistics, Law Enforcement Management And Administrative Statistics (Computer File) U.S. Department Of Commerce, Bureau Of The Census Producer, Washington, DC and Inter-university Consortium for Political and Social Research Ann Arbor, Michigan. (1992)

U.S. Department of Justice, Federal Bureau of Investigation, Crime in the United States (Computer File) (1995)

Additional references

Redmond, M. A. and A. Baveja: A Data-Driven Software Tool for Enabling Cooperative Information Sharing Among Police Departments. European Journal of Operational Research 141 (2002) 660-678.

Rousseeuw, P. J., & van Zomeren, B. C. (1990). Unmasking Multivariate Outliers and Leverage Points. Journal of the American Statistical Association, 85(411), 633–639. https://doi.org/10.1080/01621459.1990.10474920

Source - https://stackoverflow.com/a

Posted by DonJ, modified by community. See post 'Timeline' for change history