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.

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.

Methods

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.

1. To gain an initial understanding of the association between poverty rates (*PctPopUnderPov*) and violent crime rates (*ViolentCrimesPerPop*), a univariate regression is performed.
2. To analyze if the relationship is confounded by other socio-economic variables, whether there are relevant interaction effects at play and to see if the inclusion of these other socio-economic variables improves the performance of our model substantially, a multivariate regression is performed.

Before building the models, some descriptive analysis of the dataset and variables is performed. Next, the dataset is randomly split into a training dataset (80% of the data) and a holdout set (20% of the data).

For the multivariate model, an All-variable procedure is emplored to test every combination of possible variables as a model. The best model is chosen based on the Bayesian Information Criterion (BIC). Next, the assumptions of this best model are checked and partial regression plots are generated to check if each of the added variables are in the correct functional form. Afterwards, the most apropriate interaction term with the main predictor PctPopUnderPov is selected based on the same BIC procedure. Finally, multicollinearity is assessed using the Variance Inflation Factor (VIF) to obtain a more interpretable model. After determining the final model, some model diagnostics are calculated to determine outliers and influential points. Finally the model is validated by the holdout dataset.

Data preparation

The variable *NumImmig* is converted to a percentage of the total population, 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 2215) observations identified as NA values for the outcome variable *ViolentCrimesPerPop*. No imputations were used to have a more transparent model building procedure and to avoid overestimating the quality of the model fit. So these observations where not taken into account for the model building. It can be noted that the variables *countyCode* and *communityCode* are also frequently unknown.

Descriptive analysis

Univariate descriptive analysis is performed, for both the missing values and the non-missing values. Neither summary statistics nor visualisations of the predictor distributions (like boxplots and histograms) indicate that the missing data have characteristics that differ substantially from the non-missing data.

To understand how the variables in the dataset relate to one another, correlation coefficients were calculated and scatter plots were examined to visualize the relationships between the variables, the outcome variable (ViolentCrimesPerPop), and the main predictor (PctPopUnderPov).

It is notable that the variable *racepctblack* shows the strongest correlation with the outcome variable ($r = 0.63$), even surpassing *PctPopUnderPov*, the main predictor selected for this study. Given some strong correlations between predictor variables (for example between *PctNotHSGrad* and *PctLess9thGrade* ($r = 0.93$), between *perCapInc* and *PctBSorMore* ($r = 0.77$), and between *PctNotHSGrad* and *PctBSorMore* ($r = -0.75$)) and the linear relationships some variables have with *PctPopUnderPov*, it is recommended to assess multicollinearity during the model-building stage.

The scatter plots show that most variables display a roughly linear relationship with *ViolentCrimesPerPop*, although that trend is often distorted in the extreme regions of the x-axis. *agePct12t29* has a very low correlation coefficient with *ViolentCrimesPerPop* (0.11). The variable *perCapInc* shows a higher correlation (-0.32), but there is clearly no linear trend present.

Two extreme values for the *ViolentCrimesPerPop* variable appear. Chestercity has the highest number of violent crimes (4877 violent crimes per 100K population). Spencercity, on the other hand, reports zero violent crimes. Given that Spencercity is a small community (11,066 inhabitants), it is difficult to assess the accuracy of the reported value. In the model diagnostics section, we will examine whether these two observations are outliers with respect to our linear model.

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 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.

«««< HEAD

«««< HEAD Figures to get a visual illustration of whether outliers are more common in small pop. We see larger residuals for larger populations. This means that the model tends to underestimate the total number of crimes for large populations. ======= Figures to get a visual illustration of whether outliers are more common in small pop. We see larger residuals for larger populations. This means that the model tends te underestimate the total number of crimes for large populations.

=======

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:

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

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 of the 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 + 8.0268453×10^{-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 + 8.0268453×10^{-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

Model validation

Refitting the model on the test data

When the final model is fitted to the test set or the full dataset, the estimated coefficients do not differ much from those obtained when fitting the model to the training set. The more data the model is fitted on, the more significant the p-values become.

Prediction of the test data

For the multivariate model 97.2431078 % of the observed values fall within the prediction interval. This means that the prediction intervals are a bit too strict. For the univariate model 91.7293233 % of the observed values fall within the prediction interval. This means that the prediction intervals are a bit too lenient. The variance of the residuals is proportional to the magnitude of the outcome variable. The univariate model incorrectly assumes homoscedasticity in the residuals, causing prediction intervals to be too wide for low outcome values and too narrow for high outcome values. As a result, the lower bound of the prediction interval is often negative for low values. In contrast, the multivariate model (which models the log-transformed outcome variable) produces prediction intervals whose width is proportional to the magnitude of the outcome variable, which aligns better with reality. Therefore, the prediction intervals from the multivariate model are more useful than those from the univariate model.

When point estimates are compared to observed values, both the univariate and the multivariate model tends to underestimate large outcome values and tends to overestimate small outcome values. This statement holds for both the prediction of the training data and the prediction of the test data. When the mean squared prediction error (for test data) is compared with the mean squared error (for training data), it can be noticed that the predictive power of the model is better for the test data than for the training data. This indicates that there is no overfitting. The same trend is visible when the values for R^2 are compared.

The R^2 of the multivariate model is higher than the R^2 of the univariate model. However, these values are difficult to compare, because the outcome variable differs between the two models ($ViolentCrimesPerPop$ vs. $\log(ViolentCrimesPerPop)$) and back-transforming the predicted values of $\log(ViolentCrimesPerPop)$ does not yield the expected value for $ViolentCrimesPerPop$ (,except under strict assumptions regarding the residuals).

Statistical discussion

There is a large portion of missing data in our dataset (221 of the 2215 observations). The main source of missing data 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. Although the characteristics of the communities with missing data are not much different from those of the non-missing data. 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. Moreover, the LEMAS survey includes all communities with police departments of more than 100 officers, but only a random sample of smaller departments. As a result, the representation of small communities in the dataset is less precise.

Smaller communities have greater variance in both predictor and outcome variables than larger communities. «< Wat schrijven we hier nog bij?»>

Another caveat is that the FBI's data is from 1995, while the Census' populations data containing demographic variables are from 1990. If the demographic variables of certain communities changed drastically during this period, the predictor variables would not represent these communities well in 1995 (when the actual crimes happened).

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 interpreted 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.

Conclusion

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. For this purpose we used a merged dataset with detailed data on both crimes and demographic variables from 1990 and 1995.

The univariate linear regression showed a significant positive relationship between the percentage poverty and the number of crimes per 10k population. For the multivariate model, that modelled the logarithm of *ViolentCrimesPerPop*, a subset of five additional (socio-economic) predictor variables was selected. The following predictor variables show a significant positive association with the number of crimes per 10k population (*ViolentCrimesPerPop*): *PctPopUnderPov*, *race_black_logit* and *PctImmig*. The following predictor variables showed a significant negative association with the number of crimes per 10k population: *PctBSorMore* and *agePct12t29*. The selected interaction term is *PctPopUnderPov:PctBSorMore* and has a significant positive association with the number of crimes per 10k population, which means that a higher percentage of inhabitants with a bachelor degree or more implies a stronger association between poverty and the number of crimes per 10k population. The multivariate model predicts the logarithm of the number of crimes per 10k population with an R^2 of 0.5328 on the training dataset and an R^2 of 0.5841 on the test dataset. The modelling of the logarithm is very useful to give good prediction intervals, because in this way the heteroskedasticity of *ViolentCrimesPerPop* is considered, but makes the interpretation of the model less convenient and makes it difficult to deliver good point estimates for *ViolentCrimesPerPop*.

«< ik vind dit een vree grote paragraaf om aandacht te besteden aan 1 geschatte coefficient. Ik heb een stukje hierboven opgenomen.»> 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 (Kelly (2000), Fajnzylber et al. (2002) and Kang (2016)).

Several limitations should be highlighted though. The dataset is not very recent, contains missing data and uses outdated demographic predictors. Keeping this context in mind, our results show that the number of violent crimes is associated with socio-economic disadvantage. It's important to mention that all results are interpreted as correlational and not as causal. The theoretical framework, which is essential for establishing causal relationships, is not further developed in this statistical study. Moreover, the number of predictors included in the analysis is very limited, which means confounding could play a significant role.

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