Bleeding Red, White, and Blue: Predicting the Homicide Rate in the U.S. in 2019

Xinxin Li, Christopher Meng, Jessica Sang, Shikha Shrestha
Swarthmore College
April 29, 2022

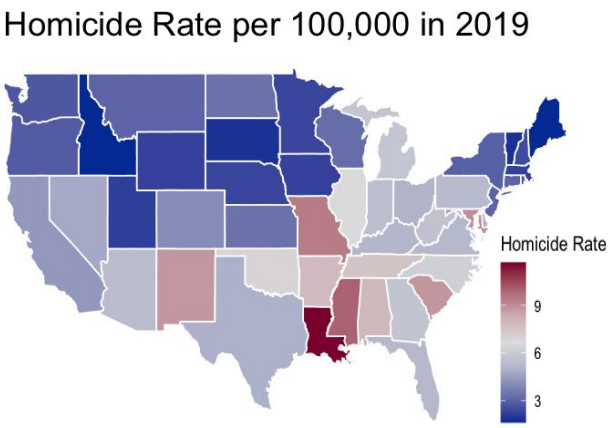
Introduction

- Crime has direct negative impacts not only on victims, but also on communities of victims
- Inspired by FiveThirtyEight's discovery on a link between hate crimes and income inequality on 2016 data (Majumder, 2017)
- Several studies have also found a link between crime/violence and income inequality, among other variables like education and social support (Hipp, 2007; Bell et al., 2018)
- With rising income inequality, what does this mean for homicides? (Horowitz et al., 2020)

Question: What variables have a relatively significant effect on the homicide rate in the U.S. in 2019?

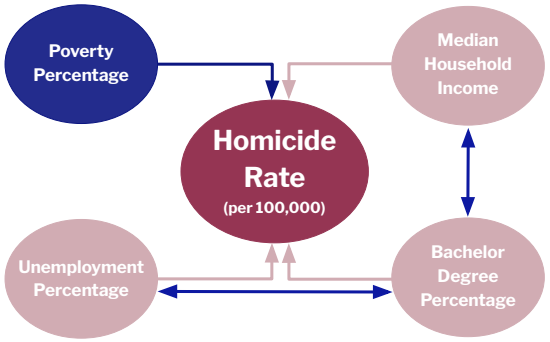
Hypothesis: We predict income inequality, as measured by the Gini index, and the poverty % to have relatively significant effects on the homicide rate.

Potential Predictor Variables (State-Level Data)	Five-Number Summary
Gini Index	0.43-0.45-0.46-0.48-0.51
Median Household Income	$44,787-$60,957-$70,031-$74,413-$95,572
Bachelor Degree Percentage	22.7%-30.1%-36.0%-40.5%-70.4%
High School Degree Percentage	86.2%-89.5%-92.1%-93.4%-97.9%
Political Side	Blue - 15, Divided - 13, Red - 23
Poverty Percentage	4.9%-8.9%-10.3%-12.4%-19.4%
Unemployment Percentage	2.1%-3.0%-3.5%-4.0%-5.5%
Uninsured Percentage	3.0%-5.9%-7.9%-10.1%-18.4%

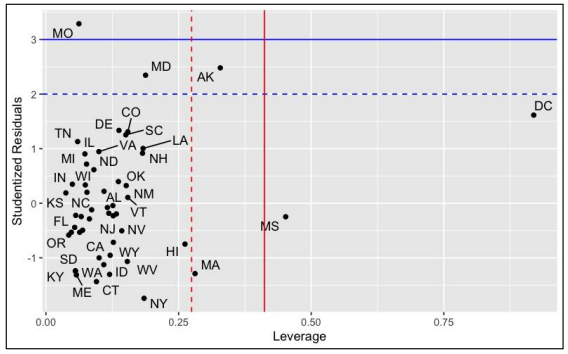


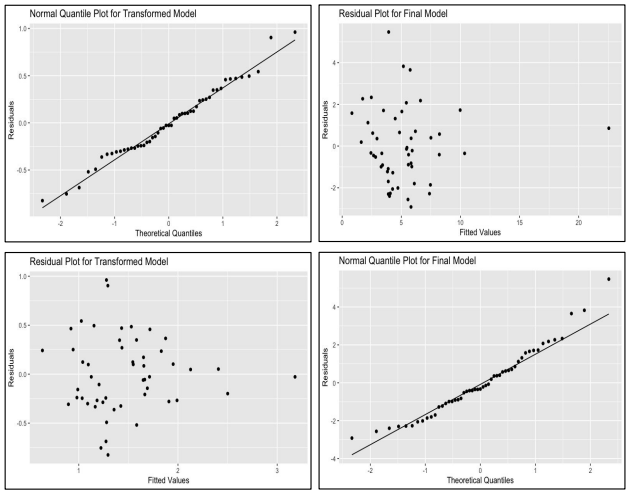
Key Takeaways

- Forward selection: poverty rate, unemployment percentage, percentage w/bachelor degree, median household income, and 2 interaction terms were chosen for our final model
  - Adjusted $R^2 = 0.7303$, $F = 23.57$ ($p < 0.001$)
- There is an association between these variables and homicide rate
- Some conditions necessary for inference were not met: linearity, constant variance
- Washington D.C. was an influential point w/extreme values for several predictors + response
  - No combination of predictors could explain >55% of variability in response w/o D.C.
- *We can't draw causal conclusions because data is observational
- *Furthermore, we have to be careful about interpreting p-values and drawing inferences because there is nothing random with the way our data was collected



Final Model Predictor Terms	Coefficients	P-values
Intercept	26.964*	0.0449
Poverty Percentage	0.696**	0.0020
Unemployment Percentage	-3.946**	0.0036
Bachelor Degree Percentage	-1.036**	0.0010
Median Household Income (1k)	-0.201	0.1752
Unemployment:Bachelor Degree	0.131***	0.0003
Median Household Income:Bachelor Degree	0.008*	0.0341





Discussion

- Ungeneralizable: cross validation w/2020 data → 56% shrinkage
- Limitations
  - Multicollinearity and interpretability
  - Violation of regression assumptions
  - Limited by time and knowledge learned in STAT021
  - FBI receives crime data from 8/10 orgs and estimates missing data
- Further research
  - Explore factors associated with homicide rate w/ county-level data
  - Examine other predictors (e.g. welfare, houselessness percentages)
    - What potential do categorical predictors have?
  - Techniques that account for violation of regression assumptions
  - How to estimate missing crime data + what is defined as crime
  - Examining how these variables are associated w/ other types of crime

References

- Bell, Brian, Rui Costa, and Stephen Machin. 2018. "Why Does Education Reduce Crime?" Center for Economic Policy Research. https://cepr.org/active/publications/discussion_papers/dp.php?dpno=13162.
- Hipp, John R. 2007. "Income Inequality, Race, and Place: Does the Distribution of Race and Class Within Neighborhoods Affect Crime Rates?" *Criminology* 45(3). https://escholarship.org/uc/item/7kw8q7hw.
- Horowitz, Juliana Menasce, Ruth Igielnik, and Rakesh Kochhar. 2020. "1. Trends in Income and Wealth Inequality." Pew Research Center's Social & Demographic Trends Project. https://www.pewresearch.org/social-trends/2020/01/09/trends-in-income-and-wealth-inequality/.
- Maiamajumder, Maimuna. "Higher Rates of Hate Crimes Are Tied to Income Inequality." FiveThirtyEight, FiveThirtyEight, 23 Jan. 2017, https://fivethirtyeight.com/features/higher-rates-of-hate-crimes-are-tied-to-income-inequality/.

CO_2rrelation between GDP, Population, and Oil Emissions

Justin Pontrella, Elizabeth Rosenthal, Satchel Tsai, & Dulce Ventura

Statistics II, Spring 2022, Presented: 4/29/2022

DATA

Source:

Our World in Data, a scientific online publication that provides free datasets on global issues like climate change

Variables:

Observational Units: One year of data from a country

Predictor Variables: GDP, Population, Carbon Dioxide Emissions from the Oil Industry

Response Variable: Carbon Dioxide Emissions

Sampling:

Conducted a stratified sampling procedure in which 10 annual observations from each country were randomly selected.

Cleaning:

Removed observations without data for predictor variables; those not classified as countries

RESEARCH QUESTION & HYPOTHESIS

Research Question:

What factors are most influential in predicting a country's total yearly carbon dioxide emissions?

Hypothesis:

A multiple regression model with all 3 predictors will likely be the best model to predict carbon dioxide emissions.

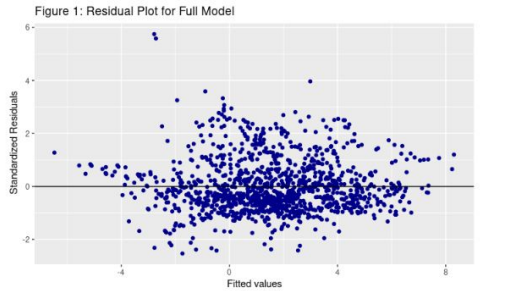
A full model that includes interaction terms to account for the potential relationships between predictors may be useful, since all of these predictors are interrelated and even dependent on each other.

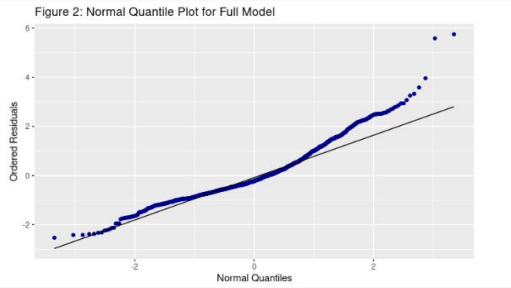
GDP will be the best predictor for a country's annual carbon emissions.

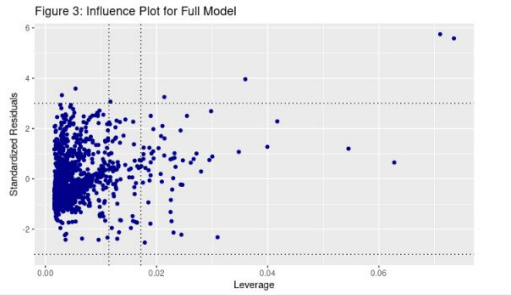
FIT & ASSESS

$\log(Y) = -23.59 + 1.04*\log(X_1) + 14.37*\log(X_2) + 1.61*\log(X_3) - 0.59*\log(X_1)*\log(X_2) - 0.05*\log(X_1)*\log(X_3) + 0.24*\log(X_2)*\log(X_3)$

X_1 = Oil CO_2 emissions, X_2 = binned population, and X_3 = GDP
Population = 1, if population is low (below a median population of 3.296 million); or 0, if otherwise







Model Output:

- Adjusted R^2 = 0.9042
- F test results:
  - p-value < 2.2e-16 ; α = 0.05
- All individual predictors were significant at α = 0.05.

Assumptions:

1. Zero mean
2. **Constant variance:** The standardized residuals vs. fitted values plot shows fairly even distribution above and below the y axis
3. Independence
4. **Linearity:** The standardized residuals vs. fitted values plot demonstrates random scattering of points.
5. **Normality:** Normal quantile plot indicates heavier tails, but not enough to violate assumption.
6. Randomness in Data Collection

Influential Points:

1. Leverage
2. Standardized Residuals
3. Cook's Distance

3 points had large residuals & leverage: Germany, 1919; NZ, 1931; and NZ, 1934.

ANOVA tests:

- Full Interaction Model vs Reduced Model (with 3 predictors)
  - p-value < 0.001; α = 0.05
- Full Interaction Model vs Reduced Model (GDP sole predictor)
  - p-value < 0.001; α = 0.05

CONCLUSION

Using the *Our World in Data* dataset, we were able to construct a multiple linear regression model that used GDP, population, carbon dioxide emissions from oil, and interaction terms that accounted for 90% of the total variation in our sample. Our analysis shows that GDP alone does a worse job of explaining the variability in our sample than models that include GDP, population, and carbon dioxide emissions.

Limitations:

Assumptions: The independence assumption was not met, as one country's GDP affects another's.

Generalizability: Due to the limitations of the data in our sample, we cannot generalize our findings beyond the period of 1868-2018. Large countries are also overrepresented in our model, due to the availability of data.

Multicollinearity: high multicollinearity means we must avoid interpreting the individual beta coefficients, preventing us from comparing the effects of each individual predictor.

REFERENCES

Dyson, Tim. 2005. "On Development, Demography and Climate Change: The End of the World as We Know It?" *Population and Environment* 27 (2): 117–49. https://doi.org/10.1007/s11111-006-0017-2.

Grossman, G. M., and A. B. Krueger. 1995. "Economic Growth and the Environment." *The Quarterly Journal of Economics* 110 (2): 353–77. https://doi.org/10.2307/2118443.

Murtaugh, Paul A., and Michael G. Schlax. 2009. "Reproduction and the Carbon Legacies of Individuals." *Global Environmental Change* 19 (1): 14–20. https://doi.org/10.1016/j.gloenvcha.2008.10.007.

Ritchie, Hannah, and Max Roser. 2020. "CO_2 and Greenhouse Gas Emissions." *Our World in Data*, May. https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions.

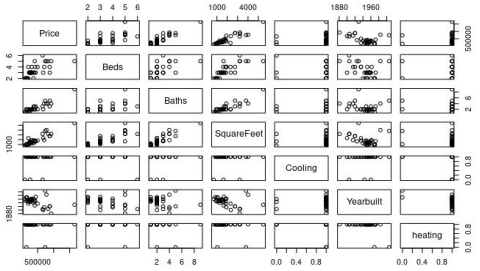
Determining the Price of a Swarthmore Dwelling: Modeling Swarthmore Housing Prices with Multiple Linear Regression

George Briggs, Elaine Kim, Max Pogorelov

April 29th, 2022

Introduction

- In 2010, about 33% of annual expenditures of homeowners was spent on housing.
- Goal: We wanted to see if we could make a model that accurately predicts house prices in Swarthmore.
- Hypothesis: There are 7 different housing variables that can be used to create a good model.
- Observational Units=Individual Houses
- 1) What variables determine house prices?
  2) Can we create an MLR model that can reasonably predict Swarthmore housing prices?

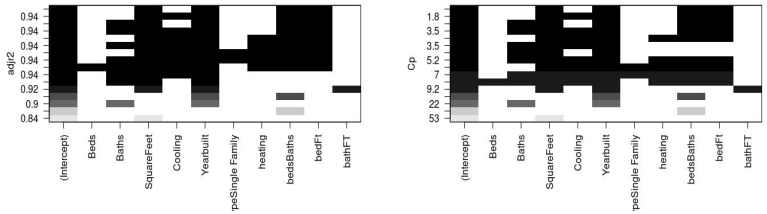


Conclusion

- The predictor variables Beds, Baths, SquareFeet, and YearBuilt were the most important for determining housing prices in Swarthmore.
- Considering its simplicity, our model did a fairly good job of predicting housing prices in Swarthmore.
- With more testing, our model could become an easy tool to roughly predict house prices.

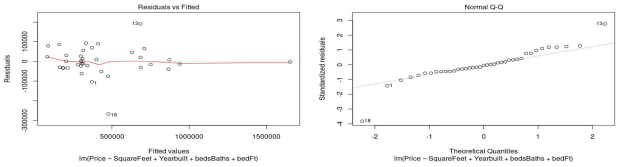
Regression Analysis

Choosing our model



$$Price_i = \beta_0 + \beta_1 SquareFeet + \beta_2 YearBuilt + \beta_3 Bedrooms*Bathrooms + \beta_4 Bedrooms*SquareFeet$$

Checking Assumptions



- Linearity
- Constant variance
- Normality
- Independence
- Randomness
- Zero mean
- No collinearity

Analyzing Model

Predictor Term	Beta coefficient Value	P-Value
Intercept	5737721.45	$2.8*10^{-6}$
SquareFeet	299.29	$2.1*10^{-5}$
YearBuilt	-2988.74	$1.9*10^{-6}$
*Bedrooms*Bathrooms*	25348.06	$6.4*10^{-6}$
*Bedrooms*SquareFeet*	-40.17	$1.6*10^{-4}$

Acknowledgements

We would like to thank Professor Suzy for teaching us these concepts and guiding us throughout our project.

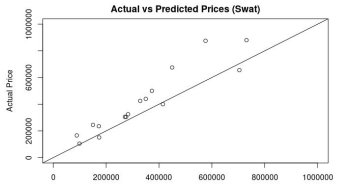
References

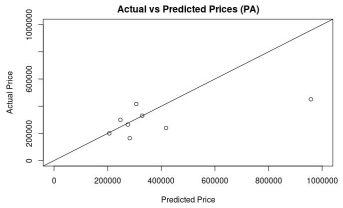
Reichenberger, Adam. n.d. "Beyond the Numbers, a Comparison of 25 Years of Consumer Expenditures by Homeowners and Renters." U.S. BUREAU OF LABOR STATISTICS
1 (15): 1–7. Accessed April 20, 2022.
https://www.bls.gov/opub/btn/volume-1/pdf/a-comparison-of-25-years-of-consumer-expenditures-by-homeowners-and-renters.pdf.
"Zillow: Real Estate, Apartments, Mortgages & Home Values." 2006-2022. 2006-2022. https://www.zillow.com. 23

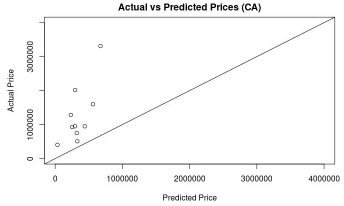
Discussion and Limitations

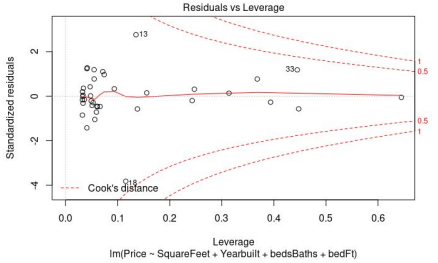
Limitations

- Cross validation results suggest that our model may not be able to accurately predict prices of houses in other locations.
- Two potential influential points, data points 13 and 18, also suggest that there is information that our model does not capture.
- Potential bias in data collection









How to Beat Basketball? Let's Ask Statistics.

April 25, 2022

Predicting NBA Win Percentage Using Regression

Brandon Cramblit, Min Nunta-Aree, Oliver Clackson

What were our goals?

- The NBA is a multi-million dollar enterprise in which teams that can gain an edge in their preparation are rewarded.
- We looked to create a model which can predict win percentage using advanced team statistics which capture more nuanced parts of the game – What leads to wins?
- In doing so, we look to make salient the elements of a basketball team which are evident of success.
- These aspects should be focused upon in future team development planning.

Where did we find our data?

- Data was acquired from basketballreference.com, which aggregates thousands of NBA statistics across time
- We randomly selected 3 teams from each NBA season from 1993 to 2019 and recorded their team statistics
- This gave us a sample of 81 teams
- This method of sampling was employed to increase independence between our datapoints
- **We used team win percentage as our response variable**
- Since certain lockout years had teams play less than 82 games, win percentage allows for integration of reduced-schedule seasons
- We looked at >50 potential predictor statistics, ranging from advanced statistics such as strength of schedule to basic statistics such as assists per game.

The Final Model

- After looking into many predictor combinations, our final model:

Win Percentage ~ True Shooting % + Turnover % + Defensive Rating + Conference

Where...

- **True Shooting %** is A measure of shooting efficiency which takes into account twos, threes, and free throws.

*TS% = Points / (2 * Field Goals Attempted + 0.44 * Free Throws attempted)*

- **Turnover %** is an estimate for the number of turnovers a team gives away per 100 possessions

*TO% = 100 * # of turnovers / (Field Goals Attempted + 0.44 * Free Throws Attempted + # of turnovers)*

- **Defensive Rating** is an estimate for the number of points allowed by a team per 100 possessions

*DRTG = (Total # points conceded (in a season) / (Field Goals Attempted - offensive rebounds + turnovers + (0.4 x Free Throws attempted))) * 100*

- **Conference** refers to the conference (West/East) that a team competes in.

Model Statistics:

- **R^2[adjusted] = 0.8584**
- **ANOVA F-test:**
  F = 115.2 ; df =4 and 76 ; p-value = **< 2.2e-16 *****
- **T-tests:**
  TS%: T =12.569 ; p-value = **< 2e-16 *****
  TO%: T =-4.889 ; p-value = **5.51e-06 *****
  DRTG: T =-17.792 ; p-value = **< 2e-16 *****
  Conference: T =2.450 ; p-value = **0.0166 ***

Other Statistics + Model Reliability

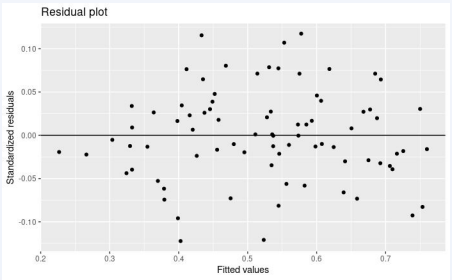
- **Mallow's Cp = 5.0**

This value is sufficient for model consideration (Cp = M + 1), and was far lower than the value for any reduced model considered.

- **Cook's D:**

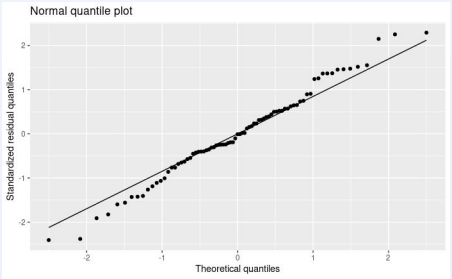
None of the residuals within the model had a greater Cook's D value than 0.5, indicating that there were no moderate to highly influential points within our dataset. – Regarding both X and Y values

Residuals vs. Fitted Values Plot



Our sample demonstrates linearity in its residuals, as well as sufficiently constant variance and a mean of 0.

Residuals vs. Normal Quantiles Plot



Our sample deviates slightly from normality, with lower-than-expected values at the lower end and higher-than-expected values at the higher end. However, this deviation is not extremely concerning.

Independence and Randomness

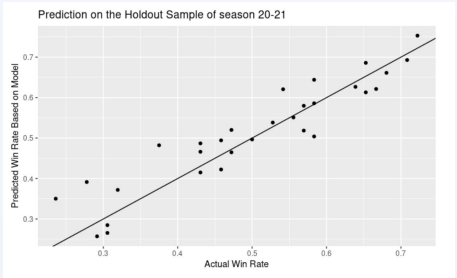
Our dataset is not entirely independent.

- Because we selected 3 teams from each season, each response value is dependent on the other values from the same season.
- Additionally, certain teams appear multiple years in a row or within 2 years of a previous appearance. These points are also dependent.

Our data was sampled randomly without replacement.

Comparison to a Holdout Sample

This training data was compared to the 20-21 NBA season, which acted as a holdout sample to test generalizability.



r = .9265

Shrinkage = -0.009562037

Our model predicts the holdout sample extremely well.

How Can This be Applied?

- In developing this model, we present possible points of emphasis for future team developments.
- By observing teams that excel in each of the listed statistics and incorporating such details into future planning, NBA teams may be provided with a source of inspiration towards success.
- Our investigation makes salient the aspects that have remained important across the tactical changes which have occurred in the NBA over the past 25 years.

Acknowledgements

We would like to thank Suzanne Thornton, as well as the entire Swarthmore Math and Statistics department.