Graduate Group Project Part II

Jessica Macchione

Sanjay Tiwari

Kendrick Turner

Jana Shepard

Melissa Kearney

Gabrielle Naimie

Tyler Ouellet

1) What do you plan to do and how will you do it?

Our plan with this graduate project is to determine if any combination of socioeconomic and household factors predict/influence crime rate. We plan to use analyses such as Multiple Linear Regression and Moderation Analysis to make these determination.

Perhaps we need to address the way we decided which tests to use to look at predictors, among other things. Perhaps looking at the research we have pulled will provide some insight.

2) What is your population?

Our population is community members varying in Age, Ethnicity, Marital Status, Education Level, and Income (to name a few) in the United States. The population we are using looks at individuals from cities such as:

* East Providence, RI
* Atlantic City, NJ
* Beverly Hills, CA
* New Haven, CT
* Lakeland, FL
* Brooklyn, NY

3) What will be your sample?

Our sample is those community members who took part in the 1990 US Census, participants of the 1990 US LEMAS Survey, and the 1995 FBI UCR.  The Sample size will be determined as per the size of expected effect, R2 = .02 (small), .13 (medium) and .26 (large). For multiple regression this works well. A general thumb rule of 10 sample per predictor will also be include as a safe rule even if the effect sizes may indicate a lower sample size.

4) Will you collect your own data (observational)?  If so what will be your variables

Our strategy is going to be using pre-existing data and doing our own analysis on it using our own choice of variables. We may do end up doing sub sampling or bootstrapping for cases when one predictor is dominating the model to reduce it’s bias.

5) Will you use a dataset from somewhere (often the internet).  If so, which one, what are your variables of interest. (must be a dataset available for public use).

We are going to be using data from the UCI Machine Learning Repository which is made accessible to the public online (<https://archive.ics.uci.edu/ml/datasets/communities+and+crime>) .

Our variables of interest include:

* Gender
* Ethnicity
* Age
* Economic Classification (Income Level)
* Single Parent Family Status
* Parental Engagement
* Population Size of Community
* Housing Status (Low Income Housing, Renting, Owning, etc)
* Average Education Level in Community (GED, High School, Associates Degree, Bachelors Degree, etc)
* Police Department Budget
* Number of Police Officers Assigned to Specialized Drug Units
* Police Officers Ethnicity
* Number of Types of Drugs Seized
* Number of Police Cars
* Average Police Officer Overtime

6) What test(s) do you think would be most appropriate?  Think about what variables you want to use.

Tests that we find most appropriate for our analysis are Multiple Linear Regression, Moderation Analysis, Levene’s Test, Ks Test, along with looking at Histograms, Frequency Distribution, PP Plots, Error and Standardized Residual, and Standardized Residual Value.

All variables for one part of the two tier analysis and then maybe the predictors we state we expect to find in question #7. Jessica

 7) What are your hypotheses, what do you think you will find and why?

Our hypotheses are as follows:

- We can predict safe communities by looking at socioeconomic factors.

- Communities with higher rates of Low-Income housing are less safe.

We believe that predictions can be made about safe communities based on socioeconomic status because communities with high poverty rates are notoriously known for having high crime rates. Examples of this include Compton, CA, Brooklyn, NY and Hartford, CT to name a few. Therefore we believe that socioeconomic status will predict if a community is safe.

- We can predict safe communities by looking at household factors.

- Communities with higher rates of Single-Parent Family Status are less safe.

We believe that family dynamics will be able to predict safety in a community. We believe that communities with high levels of Single-Parent Families, and low levels of  Parental Engagement will be considered less safe based on this research. We also assume that Housing Status (Low-Income Housing, Renting, Owning) can influence if a community is safe. For example, communities that have high rates of Low-Income Housing will be less safe than communities with high rates of Owning a home.

8) How will you be able to satisfy the conditions and assumptions for your test?

Our data Satisfies Assumptions for Regression because:

* The data comes from actual crime data, therefore the likelihood that the data is flawed is very low.
* The set up of the data has clear predictor variables and clear outcome variables.
* N>30

Assumptions of Linear Regression:

* Linear Relationship
* Normality
* No Multicollinearity
* Heterogeneity of Residuals
* No Autocorrelation
* Homoscedasticity

How to Test:

* Possible linear relationships will be established looking at a plot.
* Run a Goodness of Fit Test (Ks Test and P-Plot)
* Establish there is no Multicollinearity (Correlations Matrix, VIF, and Tolerance Statistics)
* Test for Homogeneity of Residuals using Levene’s Test
* Durbin Watson test to determine there is no Autocorrelation
* Test Homoscedasticity by plotting the data against the mean.

|  |
| --- |
| Police per population slightly positively correlated with violent crime.  B=0.258 Percentage under 9th grade positively correlated with violent crime.  B=.449 Person per family positively correlated with violent crime B=.212  Percentage unemployed positively correlated with violent crime B=.581 Household with social services positively correlated with violent crime B=.158 Female divorce positively correlated with violent crime B=.740 (HOLY CRAP) Percentage with 2 parent home negatively correlated with violent crime: B=-.818 (HOLY CRAP) Population Density positively correlated with violent crime B=.323 Median income negatively correlated with violent crime: B=-.471 Percentage under poverty line positively correlated with violent crime.  B=.532 Median rent negatively correlated with violent crime. B=-.262 Male Divorce percentage positively correlated with violent crime.  B=.671 |