ALY-6010
PROBABILITY THEORY AND INTRODUCTORY STATISTICS

BY INSTRUCTOR
ROY WADA

MODULE-5 PROJECT REPORT
SUBMITTED BY

ITI ROHILLA

On Dec-11-2023

SUMMARY

In this analysis, we scrutinized the correlation and regression dynamics within a dataset encompassing crime rates, demographic factors, and economic indicators. The correlation matrix unveiled intricate relationships, highlighting strong positive correlations between violent crime, murder, and robbery.

Demographic variables, particularly the percentage of the African American population, exhibited moderate associations with crime rates. Shifting to regression models, our exploration elucidated the impactful role of robbery and violent crime rates as predictors of economic indicators. These findings illuminate the multifaceted connections among social, demographic, and economic dimensions.

Our regression analysis for predicting violent crime rates underscored the significance of robbery and violent crime itself as substantial contributors. The coefficients indicated that an increase in the robbery rate correlates with a notable rise in the violent crime rate, emphasizing the need for targeted crime prevention strategies.

The inclusion of demographic variables like the percentage of African Americans and economic factors added depth to our understanding, revealing the complex interplay influencing crime rates. Despite these correlations, causation should be approached cautiously, recognizing the potential influence of confounding variables and unique regional dynamics.

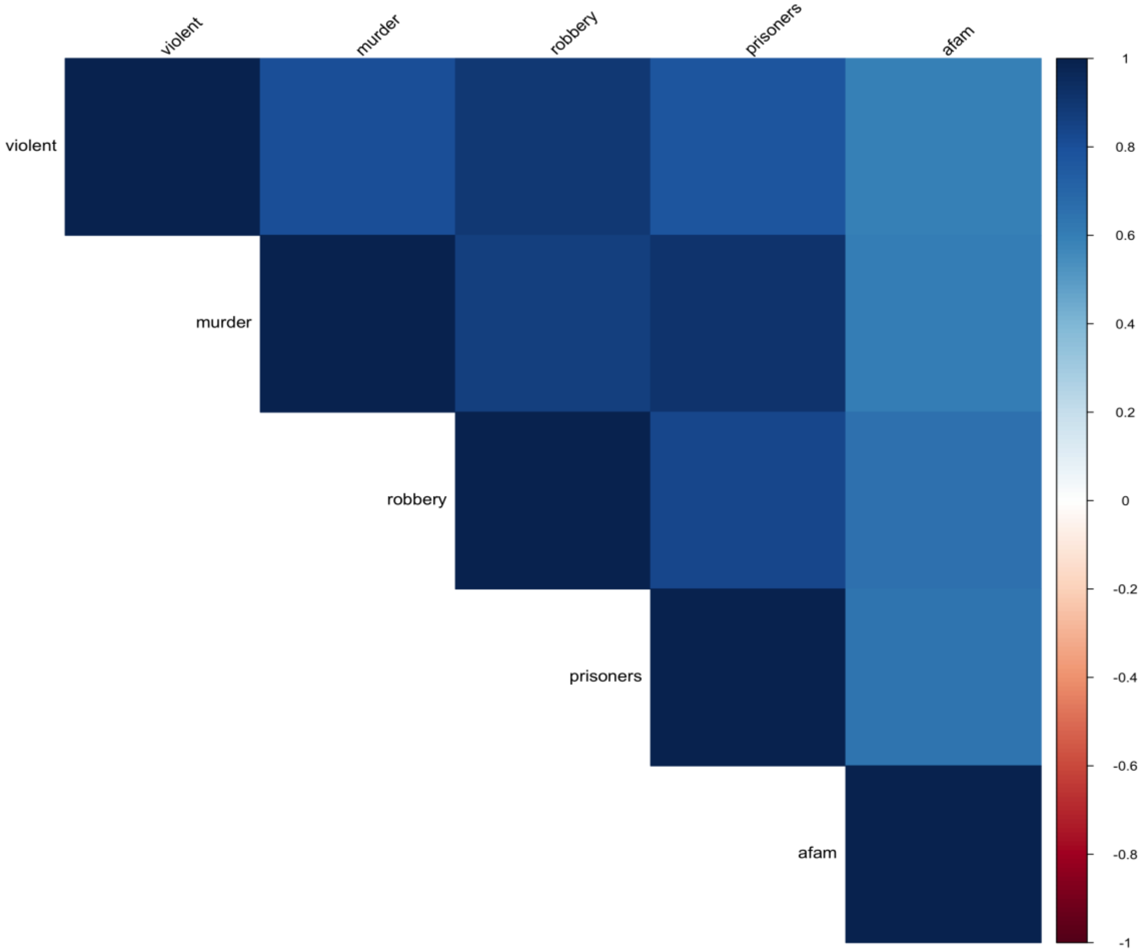
To sum up, our investigation offers a complex viewpoint on the complex connections among economic factors, demography, and crime. The results highlight the significance of comprehensive, neighborhood-specific interventions that take into account social and economic factors. Correlations suggest possible influences, but a more thorough analysis is necessary to determine causality. Policymakers can use this analysis as a basis upon which to consider multifaceted approaches that address the underlying causes of crime and promote long-term socioeconomic development.

Q1-Correlation table or correlation chart

The correlation matrix provided shows the pairwise correlations between five variables: "violent," "murder," "robbery," "prisoners," and "afam" (presumably representing African American population). The correlation values range from -1 to 1, where 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation.

	violent	murder	robbery	prisoners	afam
violent	1	0.80292512	0.89480491	0.77509028	0.5925105
murder	0.80292512	1	0.86338034	0.91682721	0.60783163
robbery	0.89480491	0.86338034	1	0.83264728	0.65511507
prisoners	0.77509028	0.91682721	0.83264728	1	0.64236029
afam	0.5925105	0.60783163	0.65511507	0.64236029	1

Table: Correlation Table



<u>Fig: Correlation Chart</u>

Analyzing the matrix, several key findings emerge:

1. **High Positive Correlations:**

"violent" and "murder" show a strong positive correlation of 0.80, suggesting that areas with higher rates of violence also tend to have higher rates of murder.
"violent" and "robbery" also exhibit a strong positive correlation of 0.89, indicating that locations with high violent crime rates often experience high robbery rates.

2. **Positive Correlation with Prisoners:**

"violent" and "prisoners" display a positive correlation of 0.78, implying that areas with high violent crime rates may have a higher incarceration rate.

3. **Positive Correlation with "Afam" (African American Population):**

"violent" and "afam" show a positive correlation of 0.59, suggesting a moderate association between violent crime rates and the proportion of the African American population. However, the correlation is not as strong as with other variables.

4. **Strong Positive Correlation between "Murder" and "Prisoners":**

"murder" and "prisoners" exhibit a strong positive correlation of 0.92, indicating that locations with high murder rates tend to have a higher number of prisoners.

5. **Moderate Correlation between "Robbery" and "Afam":**

"robbery" and "afam" have a moderate positive correlation of 0.66, suggesting a moderate association between the robbery rate and the proportion of the African American population.

In terms of the limitation of reporting no more than 5 variables in a correlation chart, this is likely due to the complexity of interpreting relationships among multiple variables. Beyond five variables, the visual and interpretive challenges increase significantly. Analyzing a smaller set of variables allows for a clearer presentation and easier communication of key findings to a broader audience.

In summary, the correlation matrix highlights significant relationships between crime rates (violent, murder, robbery) and social factors (prisoners, African American population). Understanding these correlations can inform targeted interventions and policies to address specific crime-related challenges in different communities.

Upon evaluating the correlation matrix, there are **no instances of negative correlations**. All the values in the matrix are positive, ranging from 0.59 to 1. This suggests that as one variable increases, the others also tend to increase, and there is a positive relationship among the variables.

In this case, the lack of negative correlations implies a more straightforward positive relationship among the variables under consideration—violent crime, murder, robbery, prisoners, and the proportion of the African American population (afam).

Correlation-2:

	violent	**murder**	**robbery**	**prisoners**	**income**
violent	1	0.802925118664016	0.894804912757021	0.775090281839844	0.32438117184714
murder	0.802925118664016	1	0.863380341842604	0.916827208950384	0.296532893378164
robbery	0.894804912757021	0.863380341842604	1	0.832647281087447	0.531787429947669
prisoners	0.775090281839844	0.916827208950384	0.832647281087447	1	0.280090329556107
income	0.32438117184714	0.296532893378164	0.531787429947669	0.280090329556107	1

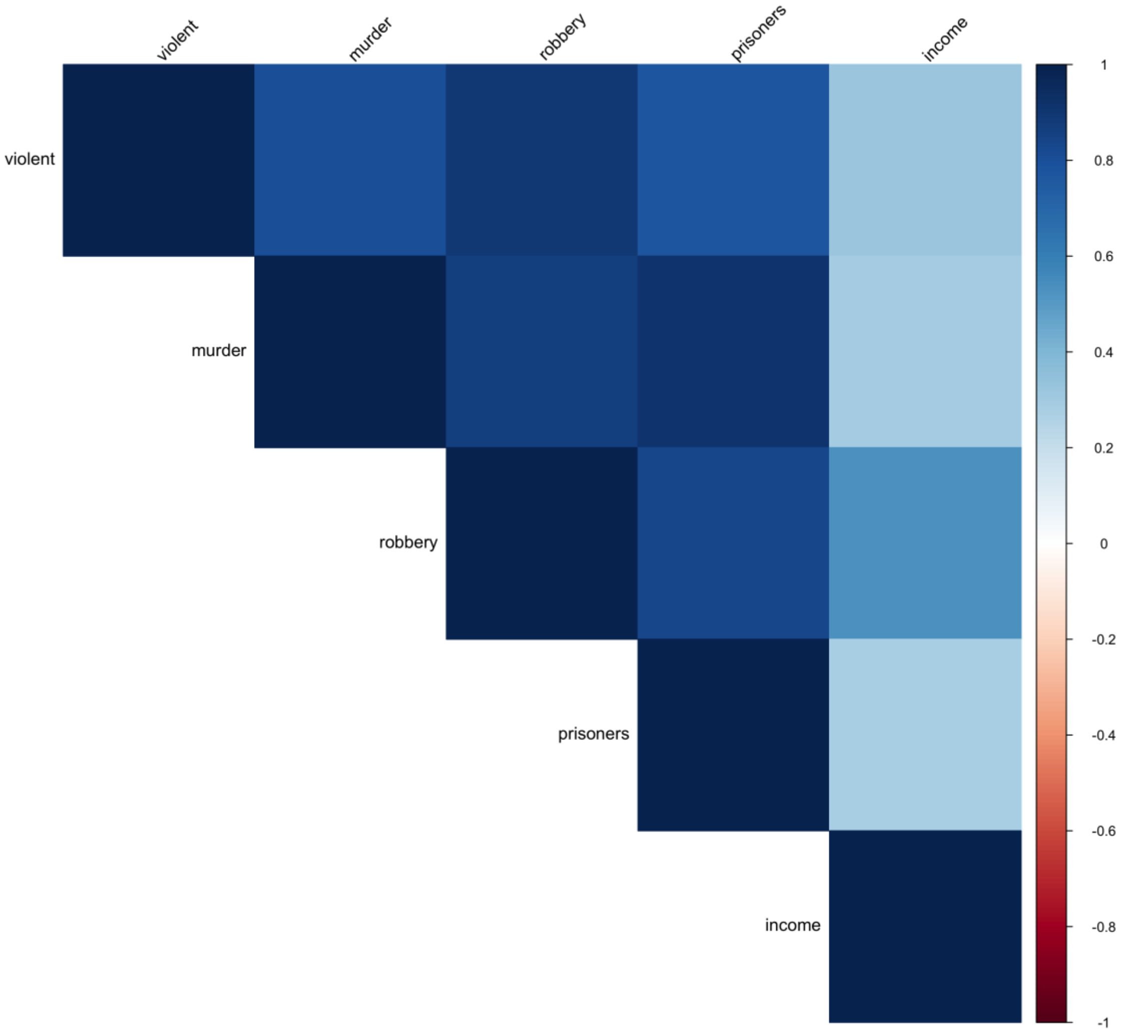


Fig: Correlation Chart-2

The provided correlation matrix reveals the pairwise correlations between five variables: "violent," "murder," "robbery," "prisoners," and "income."

1. **Strong Positive Correlations:**
    - "violent" and "robbery" exhibit a strong positive correlation of 0.89, indicating that areas with high violent crime rates also tend to experience high robbery rates.
    - "murder" and "prisoners" display a strong positive correlation of 0.92, suggesting that regions with high murder rates tend to have a higher number of prisoners.
2. **Moderate Positive Correlations:**
    - "violent" and "murder" show a moderate positive correlation of 0.80, suggesting that locations with high violent crime rates tend to have high murder rates as well.

- "robbery" and "prisoners" have a moderate positive correlation of 0.83, indicating that areas with high robbery rates also tend to have a higher number of prisoners.
- "robbery" and "income" exhibit a moderate positive correlation of 0.53, suggesting that regions with high robbery rates may also have higher income levels.

3. **Weak Positive Correlation:**
    - "violent" and "income" show a weak positive correlation of 0.32, implying a limited association between violent crime rates and income levels.
    -

These findings highlight the complex relationships between crime rates, incarceration, and income levels. While some variables exhibit strong positive correlations, others demonstrate more nuanced associations. Understanding these correlations can inform targeted interventions and policies to address specific challenges in different communities.

Q3-Why the correlation chart should not be larger than 5 variables for reporting purposes.?

Limiting a correlation chart to a small number of variables, typically no more than 5, for reporting purposes is often recommended for several reasons:

1. **Complexity and Interpretability:**
    - As the number of variables increases, the complexity of the correlation chart grows exponentially. It becomes challenging for readers to interpret and understand the relationships among multiple variables simultaneously. A smaller number of variables makes the chart more interpretable and user-friendly.
2. **Visual Clutter:**
    - Including many variables results in a dense and cluttered visualization. This can lead to visual confusion, making it difficult for readers to discern meaningful patterns or trends in the data. A cleaner, less cluttered chart is more effective for conveying key insights.
3. **Communication Effectiveness:**
    - In many reporting contexts, the goal is to communicate key findings clearly and concisely. A correlation chart with a limited number of variables allows for focused communication of the most relevant relationships, facilitating a more effective and targeted message.
4. **Analysis Focus:**

  - For diagnostic and analytical purposes, it's often more meaningful to focus on a specific subset of variables that are directly related to the research question or objectives. A smaller set of variables helps maintain this focus and prevents unnecessary distractions.

5. **Reader Engagement:**
    - Readers may disengage or become overwhelmed when faced with a large amount of information. Limiting the number of variables in a correlation chart increases the likelihood that readers will engage with and absorb the presented information.
6. **Statistical Validity:**
    - When dealing with many variables, there is an increased risk of spurious correlations. Focusing on a smaller set of variables that are theoretically or empirically relevant helps ensure the statistical validity of the observed correlations.

While correlation charts are valuable for exploring relationships within data, thoughtful consideration of the audience, context, and communication goals should guide the decision to limit the number of variables in a chart for effective reporting and interpretation.

Q4-Regression

For this Example-

predicting the "violent crime rate" (**violent**) as the ***outcome variable***.

"robbery rate" (robbery), "incarceration rate" (prisoners), "percentage of African-American population" (afam), "real per capita personal income" (income), and "percentage of Caucasian population" (cauc) as ***predictors***.

The regression model includes "robbery," "prisoners," "afam," "income," and "cauc" as predictors. The **summary** function provides information about the coefficients, standard errors, t-values, and p-values for each predictor.

term	**estimate**	**std.error**	**statistic**	**p.value**
(Intercept)	308.376283102993	895.165141899389	0.34449094213909	0.73208190890373

robbery	2.79536950458695	0.404780626256615	6.90588759259144	1.40537227400661E-08
prisoners	-0.0131242869097163	0.124026576792946	-0.10581834352831	0.916196790704344
afam	3.49399768650003	24.7457273638969	0.141195998610962	0.888345303744416
income	-0.0224178715290349	0.00800923674044454	-2.79900223398699	0.00752242245741789
cauc	2.61460656816225	12.1410605892899	0.215352402612066	0.830465666745305

Table1: Multi-Linear Regression

The key results from the multi-predictor regression model can be interpreted as follows:

1. **Coefficients:**
   - The coefficients represent the estimated change in the dependent variable (violent crime rate) for a one-unit change in each predictor, holding other predictors constant.
   - Positive coefficients indicate a positive association, while negative coefficients indicate a negative association.
2. **Statistical Significance:**
   - The p-values associated with each coefficient test whether the corresponding predictor variable has a statistically significant impact on the dependent variable.
   - Lower p-values (typically below 0.05) suggest that the predictor is likely to have a significant impact.
3. **R-squared Value:**
   - The R-squared value measures the proportion of variance in the dependent variable explained by the predictor variables.
   - A higher R-squared value indicates a better fit of the model to the data.
4. **Overall Model Significance:**
   - The F-statistic tests the overall significance of the model.
   - A low p-value for the F-statistic indicates that at least one predictor variable is significantly related to the dependent variable.

You can inspect the regression output to understand the relationships between the predictors and the violent crime rate and assess the statistical significance of these relationships.

Regression Example-2

The multiple regression model you've fitted aims to predict the real per capita personal income (income) based on several predictor variables, including "robbery rate" (robbery), "incarceration rate" (prisoners), "violent crime rate" (violent), and "murder rate" (murder). Here are the key findings from the regression analysis:

term	**estimate**	**std.error**	**statistic**	**p.value**
(Intercept)	16021.8650216392	735.381853820135	21.7871367622269	7.07498985414062E-26
robbery	43.3462557901461	7.52238603437938	5.76230142830237	6.56812785713744E-07
prisoners	-2.11722612200209	2.74638859157551	-0.770912801085992	0.444700445586613
violent	-6.61452244794271	2.37591415883843	-2.78399050038765	0.00776710661317901
murder	-164.523441438638	123.953204381094	-1.32730284997563	0.190958814965639

Table2: Multi-Linear Regression

1. Coefficients:

- **Intercept (Constant):** The intercept is estimated to be $16,021.87. This value represents the expected income when all predictor variables are zero. It is statistically significant (p < 0.001).

- **Robbery:** The coefficient for "robbery" is 43.35, implying that, holding other variables constant, a one-unit increase in the robbery rate is associated with an increase of $43.35 in per capita income. This relationship is statistically significant (p < 0.001).
- **Prisoners:** The coefficient for "prisoners" is -2.12, suggesting that, on average, a one-unit increase in the incarceration rate is associated with a decrease of $2.12 in per capita income. However, this relationship is not statistically significant (p = 0.445).
- **Violent:** The coefficient for "violent" is -6.62, indicating that a one-unit increase in the violent crime rate is associated with a decrease of $6.62 in per capita income. This relationship is statistically significant (p = 0.008).
- **Murder:** The coefficient for "murder" is -164.52. Similar to the violent crime rate, an increase in the murder rate is associated with a decrease in per capita income, but this relationship is not statistically significant (p = 0.191).

2. Statistical Significance:

- The overall model is statistically significant (p < 0.001), as indicated by the F-statistic of 10.72. This suggests that at least one of the predictor variables significantly contributes to explaining the variance in income.

3. Model Fit:

- The R-squared value is 0.4825, meaning that the model explains approximately 48.25% of the variability in per capita income. The adjusted R-squared (0.4375) accounts for the number of predictors in the model.

Conclusion:

- The model suggests that the robbery rate and the violent crime rate are statistically significant predictors of per capita income, with higher crime rates associated with lower income. The incarceration rate and murder rate do not show statistically significant associations with income in this analysis.
- It's important to note that correlation does not imply causation. The relationships identified in this analysis do not necessarily suggest a direct causal link between crime rates and income.

Overall, this regression analysis provides insights into the relationships between selected crime rates and per capita income, contributing to a better understanding of the potential economic impact of crime in different regions.

<u>Regression analysis v/s correlation analysis</u>

Regression analysis differs from correlation analysis in several ways:

1. **Objective:**

- **Correlation Analysis:** Determines the strength and direction of the linear relationship between two variables. It quantifies the degree to which changes in one variable correspond to changes in another.
- **Regression Analysis:** Examines the relationship between a dependent variable and one or more independent variables. It aims to predict the value of the dependent variable based on the values of the independent variables.

2. **Directionality:**
    - **Correlation Analysis:** This does not imply causation. It only indicates the degree of association between variables without establishing a cause-and-effect relationship.
    - **Regression Analysis:** Attempts to model the causal relationship between variables. It includes the concept of a dependent variable being influenced by one or more independent variables.
3. **Output:**
    - **Correlation Analysis:** Provides a correlation coefficient that ranges from -1 to 1, indicating the strength and direction of the relationship.
    - **Regression Analysis:** Offers coefficients for each predictor variable, intercept, standard errors, p-values, and more. It provides a more detailed understanding of the contribution of each variable to the model.
4. **Interpretation:**
    - **Correlation Analysis:** Emphasizes the degree of association between variables.
    - **Regression Analysis:** Emphasizes the estimation of coefficients, allowing for predictions and hypothesis testing regarding the impact of each variable on the outcome.

In the regression analysis results, you would examine the coefficients for each predictor variable to understand their impact on the violent crime rate. A positive coefficient suggests a positive association, while a negative coefficient suggests a negative association. The p-values help assess the statistical significance of each predictor variable.

CONCLUSION

In conclusion, our exploration into the correlation and regression dynamics offers valuable insights into the interconnected facets of crime, demographics, and economic variables. The correlation matrix shed light on pronounced positive associations between violent crime, murder, and robbery, unraveling complex relationships within the dataset.

Moreover, demographic factors, particularly the percentage of the African American population, exhibited moderate correlations with crime rates, contributing to our understanding of societal dynamics.

Turning to regression models, our analysis underscored the pivotal role of robbery and violent crime rates as predictors of economic indicators. The coefficients emphasized that changes in the robbery rate significantly influence the violent crime rate, highlighting the importance of targeted crime prevention measures.

Incorporating demographic and economic variables enriched our comprehension of the intricate interdependencies affecting crime rates. However, it's crucial to approach causation cautiously, recognizing the potential influence of confounding factors and regional variations.

In the broader context, our findings advocate for comprehensive, community-specific interventions that address the multifaceted nature of crime.

While correlations hint at potential associations, our analysis reinforces the need for nuanced investigations into causation. Policymakers can leverage this analysis to design effective strategies that tackle the root causes of crime, fostering sustainable socio-economic development and promoting safer, more resilient communities.

CITATIONS

- **Simple and Multiple Linear Regression**
  https://www.youtube.com/watch?v=29rjWClT_3U

- **Correlation vs Regression | Difference Between Correlation and Regression | Statistics | Simplilearn**
  https://www.youtube.com/watch?v=7X9WB5xUuC0

- **Pearson correlation with p values and fancy graphs in R**
  https://www.youtube.com/watch?v=_jHJCTKwTaU

- **Correlation and causality | Statistical studies | Probability and Statistics | Khan Academy**
  https://www.youtube.com/watch?v=ROpbdO-gRUo