**Economic & Well-Being Factors that Affect Per Capita Income**

Team: AB19

November 28, 2018

Ethan Jung Min Cha

Yiqi Ling

Linxi Xiao

Ayotunde Yoyin

Yichen Yuan

1. **Introduction**

As our world consists of many different regions or locations, the per capita income or PCI also varies by location. PCI is calculated by dividing a region’s total income by its population. Some areas have higher PCI's while others have much lower PCI's. Even within the same country, the PCI’s can be drastically different for each states. Besides major industries and economic activities in the region, there are many non-economic / indirect factors affecting the PCI. For example, a high work ethic in a given state can improve the productivity and thus the PCI; the perception of corruption can impact the morale of workers in a region; a high self-evaluation of life satisfaction can also have a positive effect on the economic and mental well-being (PCI) of a person. While we are not attempting to understand the causal relationships between non-economic factors and PCI, we are interested in investigating and answering the following question: Is there truly a relationship between well-being related non-economic factors and income per capita in a region?

In this report, we define well-being to be the state of being happy, healthy, or prosperous¹ as given by the Merriam-Webster dictionary and economic well-being as having present and future financial security[²](http://graphemica.com/%C2%B2) defined by the Council on Social Work Education. The general belief about well-being is that it is an important aspect of life to maintain because it is a vital criteria in measuring the overall health of an individual. A highly positive well-being is able to affect the PCI of an individual in a positive way. An excerpt from a *Huffington Post* article titled “Why Happy People Earn More Money” states that people who are happier (having a more positive well-being) tend to take more optimistic approaches to situations in life. Lynda Spiegel, a human resources professional says that employees who are happy are more open to opportunity and new experiences in life. Spiegel says “they are willing to accept challenges and take risks, both of which are predictors of greater earning opportunity”[³](http://graphemica.com/%C2%B3). The analysis, done in this report, has the ability to support the argument made by Spiegel in determining how well-being affects a country’s economic status.

1. **Problem Statement**

This project focused on data collected in the year 2014. Our aim was to give a relative perspective on what environmental factors are the main factors affecting a region's per capita income. The goal of this analysis was to determine correlations (if any) between a region's environmental status and PCI values to have a better understanding of what phenomena occur in tandem within each region selected. Looking at these similarities and differences, we are then able to show how much the general well-being of an area affects the affluence of a given area. At the end of this report, we will be able to recommend which variables should be further explored in order to increase the PCI of a region.

The Organisation for Economic Co-operation and Development (OECD), conducts its own surveys and accumulates information from government agencies from different countries in order to measure each country member’s income, poverty level, education, wellness, etc. We will focus on income data and attempt to determine which economic factors and well-being factors correlate to a region’s income level through regression analysis and modeling. We will first identify key significant variables that attribute to high economic status in terms of income per capita. With these variables, we will then do further analysis in determining how viable the given model is and whether or not it requires adjustment.

Within the scope of our research, many factors of well-being can be accounted for and may make for a useful model. Factors dealing with certain demographics such as number of residents per household, number of rooms shared per person in a household, and the employment rate of an area may be strong predictors in determining the main influencers of the PCI of the given area. For example, it could be the case that the more individuals live in a house and share more rooms, the less financially independent they are. In addition to common economic and wellness factors above, we also seek to consider other variables, such as air pollution, voter turnout rate, and perception of corruption, in order to bring new perspective on seemingly irrelevant yet influential factors on a region’s and eventually country’s income per capita. For instance, we suspect that there is a correlation in a country’s wealth being concentrated into a small proportion of individuals at a country with high corruption. We are curious to learn more about unrealized correlations between other qualitative variables and income.

By building a model that identifies significant attributes that contribute to a OECD member country’s economic prosperity, we hope to respond with a recommendation that identifies and addresses the current roadblocks into a nation’s direction toward economic prosperity. We will fit our data onto a first-order model through a careful screening and diagnostic processes and, if needed, conduct model transformations and/or construct second-order models in order to reap more purposeful conclusions.

1. **Data Description**
2. Data collection

**Our data were primarily extracted from different databases on OECD Regional Statistics. A total of 363 data points and 15 independent variables were evaluated in this study.**

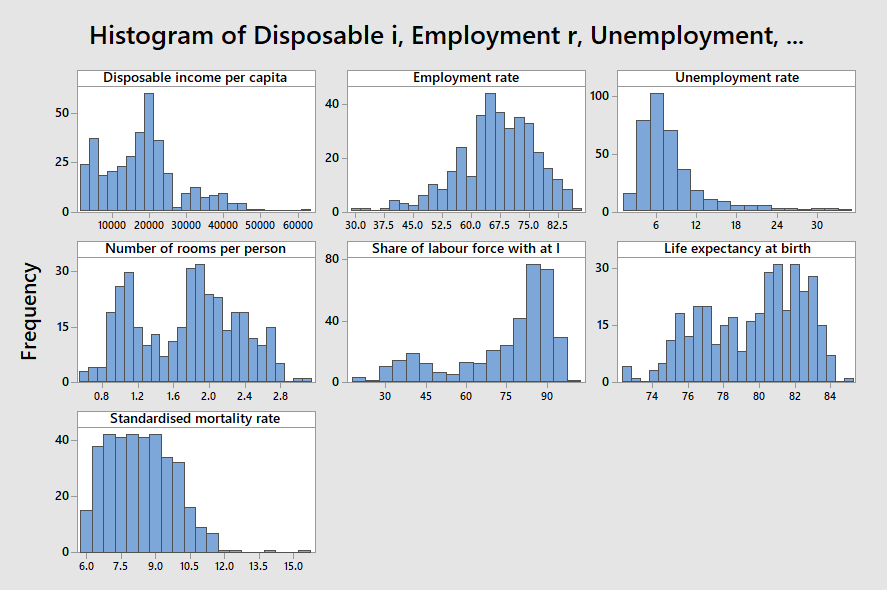
There are eleven dimensions that can be used to measure the well-being of a specific region based on OECD’s publication *How’s Life in Your Region*. We chose **nine** of them to analyze in our study, which are **housing, jobs, health status, environmental quality, security, political involvement, social connection, accessibility of services, and subjective well-being**,because we intuitively considered them to be stronger influencers. Among the nine dimensions that we are interested in investigating, we chose specific indicators to represent those categories. For jobs, we chose *employment* and *unemployment rate* as the indicators because they are the direct representation of how well people can meet their living needs. The corresponding data were extracted from the OECD Regional Labour database. To measure housing situation, we chose *number of rooms per person* as the primary indicator because it provides insights into how much living space each person gets to own. Our primary data reference comes from the OECD regional housing indicator database. We chose *life expectancy at birth* and *standardized mortality rate* as indicators for health status because they not only represent a region’s general living condition, but also can infer its medical standards. The data for these two indicators were collected and cleansed from OECD Regional demography: Mortality and Life Expectancy. Environmental quality is measured by *PM 2.5*, the most common indicator for air pollution. We simply extracted those data from the OECD Regional Environmental indicators database. Security, political involvement, social connection, accessibility of services and subjective well-being are measured by *homicide rate*, Voter turnout in general election & perception of corruption index *perceived social network support*, *share of household with internet broadband access* and *self-evaluation of life satisfaction*. Every data point was extracted from the OECD Regional social indicators database. Finally, we added *continent* as our regional indicator because it effectively generalizes all the regions we investigated. The continent data were created and processed manually by categorizing each region into its corresponding continent.

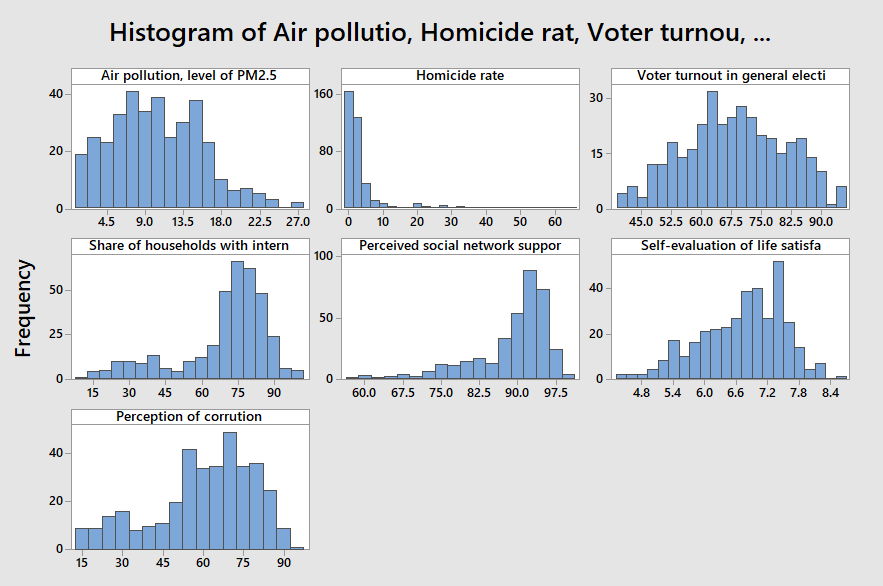
1. Data description

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variables** | **Descriptions** | **Units** | **Stats** | **Type** |
| Disposable income per capita | The average disposable income per year for each person | US Dollar | Mean: 18290.33  Range: (1786, 61738)  **Variance:** | Response;  Quantitative |
| Employment Rate | The employment rate at each region | Percentage | Mean:66.31  Range: (29.6, 86.5) | Independent;  Quantitative |
| Unemployment Rate | The unemployment rate at each region. The natural rate of unemployment is between 4.5 - 5.0 percent. | Percentage | Mean: 8.16  Range: (1.5, 34.8) | Independent;  Quantitative |
| Number of rooms per person | The average number of rooms shared per person | Ratio | Mean: 1.73  Range: (0.6, 3.1) | Independent;  Quantitative |
| Share of labour force with at least secondary education | The percentage share of labor force of whom have completed at least secondary education | Percentage | Mean: 74.96  Range: (19.9, 97.7) | Independent;  Quantitative |
| Life expectancy at birth | How long, on average, a newborn can expect to live | Years | Mean: 79.64  Range: (71.3, 84.8) | Independent;  Quantitative |
| Standardized mortality rate | A weighted average of the age-specific death rates of a given population | Per 1000 inhabitants | Mean: 8.34  Range: (5.8, 14.5) | Independent;  Quantitative |
| Air pollution, level of PM2.5 | Fine inhalable particles, with diameters that are generally 2.5 micrometers and smaller. Usually a PM2.5 level of above 35.5 is considered as unhealthy. | Micrograms per cubic metre | Mean: 10.64  Range: (0.4, 27) | Independent;  Quantitative |
| Homicide rate | The rate of unlawful death purposefully inflicted on a person by another person | Per 100000 inhabitants | Mean: 3.41  Range: (0, 64.8) | Independent;  Quantitative |
| Voter turnout in general election | The percentage of eligible voters who cast a ballot in an election | Percentage | Mean: 67.96  Range: (38.8, 95.1) | Independent;  Quantitative |
| Share of households with internet broadband access | The percentage share of households that have internet broadband access | Percentage | Mean: 69.75  Range: (9.5, 99.2) | Independent;  Quantitative |
| Perceived social network support | The percentage of people who report having relatives or friends they can count on | Percentage | Mean: 88.90  Range: (58.4, 100) | Independent;  Quantitative |
| Self-evaluation of life satisfaction | The self-evaluation of life satisfaction on scale 1-10 | Index | Mean: 6.65  Range: (4.4, 8.6) | Independent;  Quantitative |
| Perception of corruption | The corruption perception index scores countries on how corrupt their countries are believed to be. | Percentage | Mean: 61.08  Range: (13.4, 93.3) | Independent;  Quantitative |
| Continent | The continent that a country belongs to | Asia, Oceania, North America, South America, Europe | Binary variable:  (0,1) | Independent;  Categorical;  Qualitative |

1. Basic Analysis

**The two figures below provide histograms for all quantitative variables, capturing their shapes and types and distribution. About half of them follows normal distribution while the other half follows skewed distribution.**





**Histograms of all quantitative variables**

The matrix plot of all quantitative variables and their correlation matrix (in Appendix A) demonstrates potential correlations between each pair. Some of the variable pairs indicate strong positive/negative correlation. For example, *employment* vs. *unemployment*, *life expectancy at birth* vs *standardized mortality rate*, etc. Methods of dealing with correlated independent variables will be further investigated in the second-order modeling.

1. **First-Order Model**
2. Model Screening Methods

The very first step to grasp the general understanding of the relationship between independent variables and the dependent variable (Disposable income per capita) was to perform a simple linear regression, stepwise regression, forward elimination regression and lastly backward elimination regression.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Regression Method** | **# of Ind. Var** | **R2 (adj)** | **Assumption** | | |
| **Multicollinearity**  ***(# of variables with VIF > 10)*** | **Constant Variance**  ***(Residuals vs. Fitted Values)*** | **Normality**  ***(Anderson-Darling Test)*** |
| Simple Linear Regression | 17  (out of 17) | 80.80% | Not satisfied  1 variable  *(Life expectancy at birth)* | Not satisfied  *(fans outward)* | Not satisfied  *(p-value < 0.005)* |
| Stepwise  *(entry and removal*  *alpha = 0.1)* | 11  (out of 17) | 81.29% | Satisfied  0 variables | Not satisfied  *(fans outward)* | Not satisfied  *(p-value < 0.005)* |
| Forward Elimination  *(alpha = 0.1)* | 11  (out of 17) | 81.29% | Satisfied  0 variables | Not satisfied  *(fans outward)* | Not satisfied  *(p-value < 0.005)* |
| Backward Elimination\*  *(alpha = 0.1)* | 11  (out of 17) | 81.36% | Satisfied  0 variables | Not satisfied  *(fans outward)* | Not satisfied  *(p-value < 0.005)* |

Actual output in APPENDIX 6.b.1

‘\*’ indicates best first-order model

*Insignificant Independent Variables*

At significant level of 0.05, simple linear regression model had a total 9 insignificant independent variables out of 17 independent variables. Those include: Employment Rate, Unemployment Rate, Life expectancy at birth, Standardized Mortality Rate, Air pollution level of PM 2.5, Homicide Rate, and Perceived social network support, Continent\_Europe and Continent\_Oceania.

On the other hand, due iterative process of variable selection, the iterative regression methods (Stepwise, Forward Elimination, and Backward Elimination) had a total of 2 insignificant independent variables out of 11 independent variables: Continent\_Europe and Continent\_Oceania.

*Multicollinearity*

In order to identify independent variables correlated with other independent variables, we took a look at variables with variance inflation factor (VIF) above 10, an appropriate indicator of the presence of multicollinearity. In the simple linear regression model, VIF of independent variable (Life expectancy at birth) was 10.70, slightly exceeding 10. Through the matrix plot, we were able to identify that independent variables (Life expectancy at birth and Standardized Mortality Rate) were negatively correlated.

|  |
| --- |
|  |
| **Scatterplot of**  **Standardized Mortality Rate vs. Life Expectancy at Birth** |

*Identical Distribution with Constant Variance*

One of the very first residual assumptions we see is the “i.i.d.”. The assumption states that the residuals should be identically and independently distributed. Through the Residual vs. Fitted Value plot, we can determine if identical distribution assumption is satisfied. Unfortunately, all regression methodologies failed to satisfy the constant variance assumption due to the fanning out pattern visible in each regression model’s Residual vs. Fitted Value plot.

|  |
| --- |
|  |
| **Residual vs. Fitted Value plot**  **for all regression methodologies listed above** |

*Normality of Residuals*

The very next residual assumption we is the N(0, σ2). The assumption states that the residuals should be normally distributed with a mean of 0 and constant variance. Through normality plot and Anderson-Darling (A-D) test, we are able to notice:

1. Due to very low p-value (<0.005), we reject A-D test’s H0: Residuals are normally distributed.

2. The mean is 9.42E-13, around zero.

|  |
| --- |
|  |
| **Normality plot with A-D Test** |

*Influential Points (Outliers and Unusual Observations)*

As visible in the APPENDIX 6.b.1, we were able to identify 27 influential points, 6 of which are unusual observations and the remaining 21 data points are outliers. More information on influential points and handling influential points will be discussed later.

*Conclusion*

In conclusion, to be able to address the violation on multicollinearity and the fanning out shape of Residual vs. Fitted Values plot, we decided to, first, remove one independent variable with VIF greater than 10 (Life expectancy at birth) and perform a natural logarithm transformation on the dependent variable.

1. Dependent Variable Transformation (Natural Logarithm Transformation)

Given the output of the regression methods and the analyses, especially on the fanning out pattern on Residual vs. Fitted Values plot, we concluded to perform a natural log transformation on the dependent variable. After having transformed the dependent variable, we witnessed a significant increase in R2 (adjusted) from 81.36% of the Stepwise Regression before transformation to 91.37% of the simple linear regression model with natural log (ln) transformed dependent variable. Despite the increase in the percentage of data points explained by the transformed regression model, there were still violated assumptions, while multicollinearity remaining constant.

*Identical Distribution with Constant Variance*

We can see a big improvement in randomness and spread throughout the Residual vs. Fitted Value plot. However, the data points on the right side are more concentrated than those on the left side.

|  |  |  |
| --- | --- | --- |
|  | **→** |  |
| **Residual vs. Fitted Value plot**  **before Ln Transformation** |  | **Residual vs. Fitted Value plot**  **after Ln Transformation** |

*Normality of Residuals*

Once again, normality assumption has not been satisfied after LN transformation of the dependent variable. However, we can see improvements in the p-value from < 0.005 before transformation to 0.022 after transformation. Additionally, even though not visible in this graph, we identified outliers with large residuals that attribute to violation of normality assumption.

|  |  |  |
| --- | --- | --- |
|  | **→** |  |
| **Normality plot with A-D Test**  **before Ln Transformation** |  | **Normality plot with A-D Test**  **after Ln Transformation** |

*Influential Points (Outliers and Unusual Observations)*

As visible in the APPENDIX 6.b.2, we were able to identify 28 influential points, 6 of which are unusual observations and the remaining 22 data points are outliers. More information on influential points and handling influential points will be discussed later.

1. Model Screening after Dependent Variable Transformation

After having transformed the dependent variable, we performed the model screening process again with the following changes mentioned above. This time, we undertook two screening methods (simple linear regression with manual variable selection and stepwise regression), for forward elimination and backward elimination resulted in a lower R2 (adjusted) with equal or more assumption violations.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Regression Method** | **# of Ind. Var** | **R2 (adj)** | **Assumption** | | |
| **Multicollinearity**  ***(# of variables with VIF > 10)*** | **Constant Variance**  ***(Residuals vs. Fitted Values)*** | **Normality**  ***(Anderson-Darling Test)*** |
| Simple Linear Regression  (with manual variable selection) | 12  (out of 16) | 91.39% | Satisfied  0 variables | Satisfied | Satisfied |
| Stepwise  *(entry and removal*  *alpha = 0.1)* | 12  (out of 16) | 91.39% | Satisfied  0 variables | Satisfied | Satisfied |

*Insignificant Independent Variables*

In Simple Linear Regression model with manual variable selection, we removed 4 insignificant independent variables (Unemployment Rate with p-value of 0.386, Air Pollution with p-value of 0.533, Perceived Social Network Support with p-value of 0.144, and Self-evaluation of Life Satisfaction with p-value of 0.588) at significance level of 0.10. After removal, the output of the model was identical of stepwise regression model. After having removed insignificant independent variables, R2 (adjusted) increased by 0.02% from 91.37% to 91.39%.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Coefficients   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Term | Coef | SE Coef | T-Value | P-Value | VIF | | Constant | 6.839 | 0.336 | 20.38 | 0.000 |  | | Employment rate | -0.00405 | 0.00229 | -1.77 | 0.077 | 4.22 | | Unemployment rate | -0.00348 | 0.00401 | -0.87 | 0.386 | 3.55 | | Number of rooms per person | 0.6169 | 0.0392 | 15.73 | 0.000 | 4.06 | | Share of labour force with at l | 0.00602 | 0.00120 | 5.00 | 0.000 | 4.38 | | Standardised mortality rate | -0.0335 | 0.0102 | -3.30 | 0.001 | 1.80 | | Air pollution, level of PM2.5 | -0.00191 | 0.00307 | -0.62 | 0.533 | 2.17 | | Homicide rate | -0.00693 | 0.00236 | -2.94 | 0.004 | 2.07 | | Voter turnout in general electi | 0.00379 | 0.00141 | 2.68 | 0.008 | 2.59 | | Share of households with intern | 0.01515 | 0.00133 | 11.36 | 0.000 | 5.14 | | Perceived social network suppor | 0.00343 | 0.00234 | 1.46 | 0.144 | 2.58 | | Self-evaluation of life satisfa | -0.0131 | 0.0242 | -0.54 | 0.588 | 2.91 | | Perception of corruption | 0.004147 | 0.000870 | 4.77 | 0.000 | 2.16 | | Continent |  |  |  |  |  | | Asia | 0.000000 | 0.000000 | \* | \* | \* | | Europe | 0.0482 | 0.0464 | 1.04 | 0.299 | 4.29 | | North America | 0.1336 | 0.0659 | 2.03 | 0.043 | 6.61 | | Oceania | 0.0411 | 0.0776 | 0.53 | 0.597 | 2.73 | | South America | 0.2974 | 0.0790 | 3.77 | 0.000 | 2.59 | | **→** | |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Coefficients  Term | Coef | SE Coef | T-Value | P-Value | VIF | | Constant | 6.870 | 0.212 | 32.46 | 0.000 |  | | Employment rate | -0.00296 | 0.00160 | -1.85 | 0.065 | 2.06 | | Number of rooms per person | 0.6298 | 0.0346 | 18.21 | 0.000 | 3.16 | | Share of labour force with at l | 0.00646 | 0.00109 | 5.91 | 0.000 | 3.62 | | Standardised mortality rate | -0.03147 | 0.00958 | -3.28 | 0.001 | 1.60 | | Homicide rate | -0.00690 | 0.00233 | -2.96 | 0.003 | 2.03 | | Voter turnout in general electi | 0.00376 | 0.00130 | 2.90 | 0.004 | 2.18 | | Share of households with intern | 0.01506 | 0.00125 | 12.01 | 0.000 | 4.55 | | Perception of corruption | 0.004038 | 0.000796 | 5.07 | 0.000 | 1.81 | | Continent |  |  |  |  |  | | Asia | 0.000000 | 0.000000 | \* | \* | \* | | Europe | 0.0577 | 0.0391 | 1.48 | 0.140 | 3.05 | | North America | 0.1362 | 0.0542 | 2.51 | 0.012 | 4.47 | | Oceania | 0.0608 | 0.0673 | 0.90 | 0.366 | 2.06 | | South America | 0.2960 | 0.0734 | 4.03 | 0.000 | 2.24 | |
| **Coefficients table**  **before variable selection** |  | **Coefficients table**  **after variable selection** |

*Identical Distribution with Constant Variance*

We can see a small improvement in randomness and spread throughout the Residual vs. Fitted Value plot. However, the data points on the right side are still more concentrated than those on the left side.

|  |  |  |
| --- | --- | --- |
|  | **→** |  |
| **Residual vs. Fitted Value plot**  **before variable selection** |  | **Residual vs. Fitted Value plot**  **after variable selection** |

*Normality of Residuals*

Once again, normality assumption has not been satisfied after the variable selection process. Surprisingly, there has been a decrease in p-value by 0.001 from p-value of 0.022 to 0.021. As a result, in order to address this issue, we focused on handling influential observations and unusual observations to assure normality assumption.

|  |  |  |
| --- | --- | --- |
|  | **→** |  |
| **Normality plot with A-D Test**  **before variable selection** |  | **Normality plot with A-D Test**  **after variable selection** |

*Influential Points (Outliers and Unusual Observations)*

Instead of using the “Fits and Diagnostics for Unusual Observations” table in the output after variable selection, we identified 18 observation points (9 of which were outliers with |SRES| and |TRES| > 2.5 and the remaining 9 were unusual observations with Cook’s Distance exceeding 0.01143). We will discuss handling influential points in the next part.

1. Handling Influential Points

**We used the rules of thumb:**

1. |SRES| and |TRES| > 2.5 (instead of 2) to identify outliers

2. Cook’s Distance: Di > 4/(n-(k+1)) = 0.01143 to identify unusual observations.

Many of the identified outliers and high leverage points identified with hii > 2(k+1)/n fell under data points marked as unusual observations through Cook’s Distance. As a result, through this process, we identified 18 significant outliers and unusual observations.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | **Region** | **Country** | **Continent** | | TRA1: Northeastern Anatolia - West | TUR | Asia | | EE007: Northeast Estonia | EST | Europe | | AU8: Australian Capital Territory (ACT) | AUS | Oceania | | CA10: Newfoundland and Labrador | CAN | North America | | ME29: Tlaxcala | MEX | North America | | CA11: Prince Edward Island | CAN | North America | | PT30: Madeira (PT) | PRT | Europe | | CA13: New Brunswick | CAN | North America | | CL11: Aysén | CHE | South America | | CA12: Nova Scotia | CAN | North America | | US11: District of Columbia | USA | North America | | US02: Alaska | USA | North America | | IL06: Southern District | ISR | Asia | | AU7: Northern Territory (NT) | AUS | Oceania | | US15: Hawaii | USA | North America | | SK01: Bratislava Region | SVK | Europe | | PT20: Azores (PT) | PRT | Europe | | US06: California | USA | North America |   **Influential and Unusual Observations**  **Identified and Removed** |

**After having removed 18 observations, we witnessed increase in R2 (adjusted) from 91.39% before removal of observations to 93.96% after removal of observations.**

*Identical Distribution with Constant Variance*

We can see a improvement in constant variance assumption. Before, the range of residuals was between -0.5 and 0.8. Now, the new range of residuals fall between -0.5 and 0.5, demonstrating roughly equal number and range of positive and negative residual observations.

|  |  |  |
| --- | --- | --- |
|  | **→** |  |
| **Residual vs. Fitted Value plot**  **before observation removal** |  | **Residual vs. Fitted Value plot**  **after observation removal** |

*Normality of Residuals*

Here, we witness satisfaction of the residual normality assumption with a significant increase in p-value of 0.983. As a result, it fails to reject A-D test’s H0: Residuals are normally distributed.

|  |  |  |
| --- | --- | --- |
|  | **→** |  |
| **Normality plot with A-D Test**  **before observation removal** |  | **Normality plot with A-D Test**  **after observation removal** |

*Conclusion*

In conclusion, removing influential and unusual observations, we were not only able to increase the percentage of observations explained by the model but also improve residual assumptions that weren’t satisfied before removing observations.

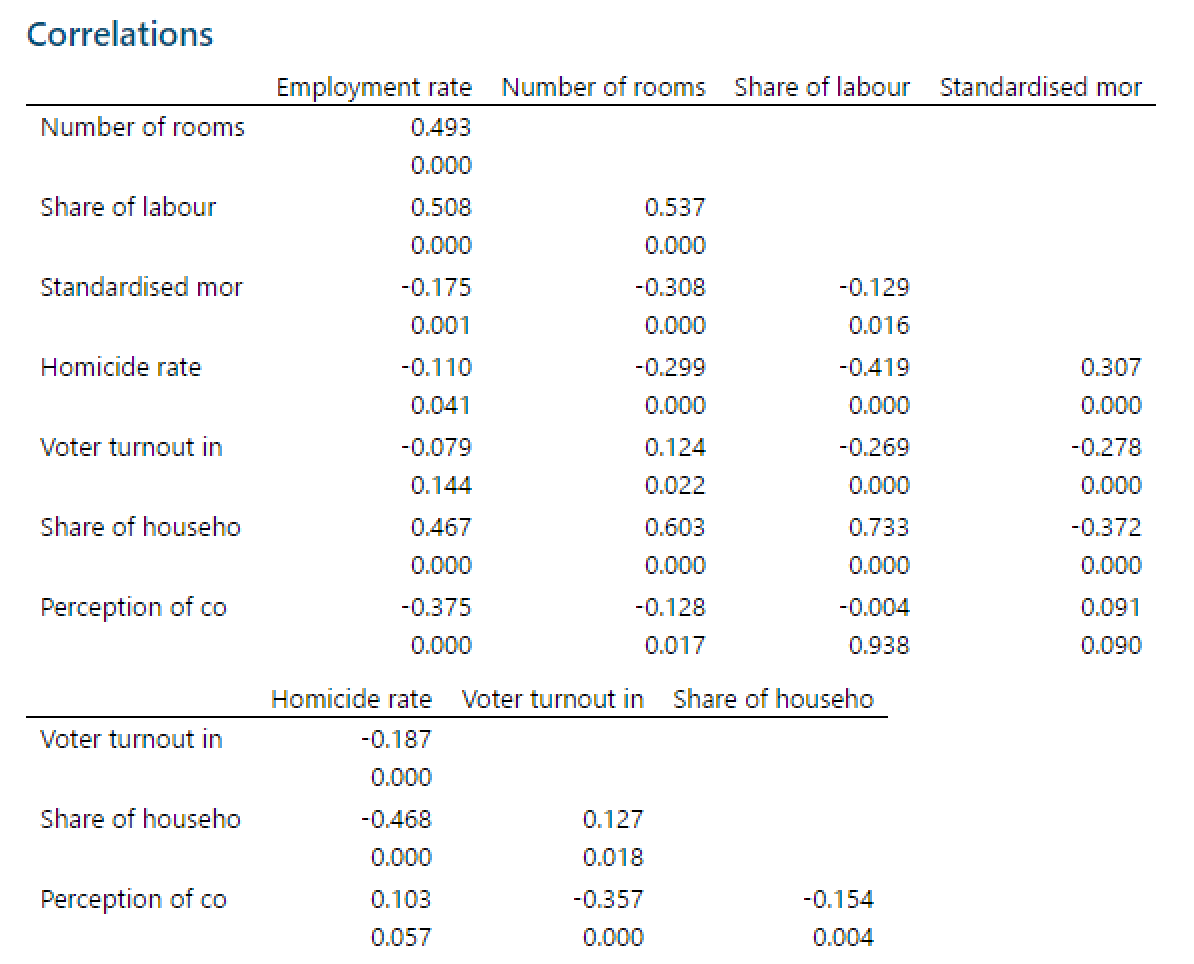
**First-order Final Model**

Regression Equation:

|  |  |  |
| --- | --- | --- |
| ln y | = | 7.050 - 0.00424 Employment rate + 0.6433 Number of rooms per person  + 0.007337 Share of labour force with at l - 0.03840 Standardized mortality rate  - 0.00774 Homicide rate + 0.00381 Voter turnout in general electi  + 0.01407 Share of households with intern + 0.003126 Perception of corruption  + 0.0 Continent\_Asia + 0.0494 Continent\_Europe + 0.1445 Continent\_North America  - 0.0185 Continent\_Oceania + 0.2439 Continent\_South America |

1. **Second-Order Model**

As noticed in the residual plot of the first-order model, though the constant variance assumption had some improvement after removing unusual observations, the residuals are still slightly fanning out. In order to further improve the constant variance assumption of the first-order model, we decided to include second order terms in our model. First, we observe the residual versus independent variables plots and there is no parabolic or curved patterns. Therefore, we decided not to add quadratic terms in the model. Next, from the correlation matrix shown below, we have identified some correlation between several independent variables. It is appropriate to add interaction terms if the value of one independent variable is dependent on the other.

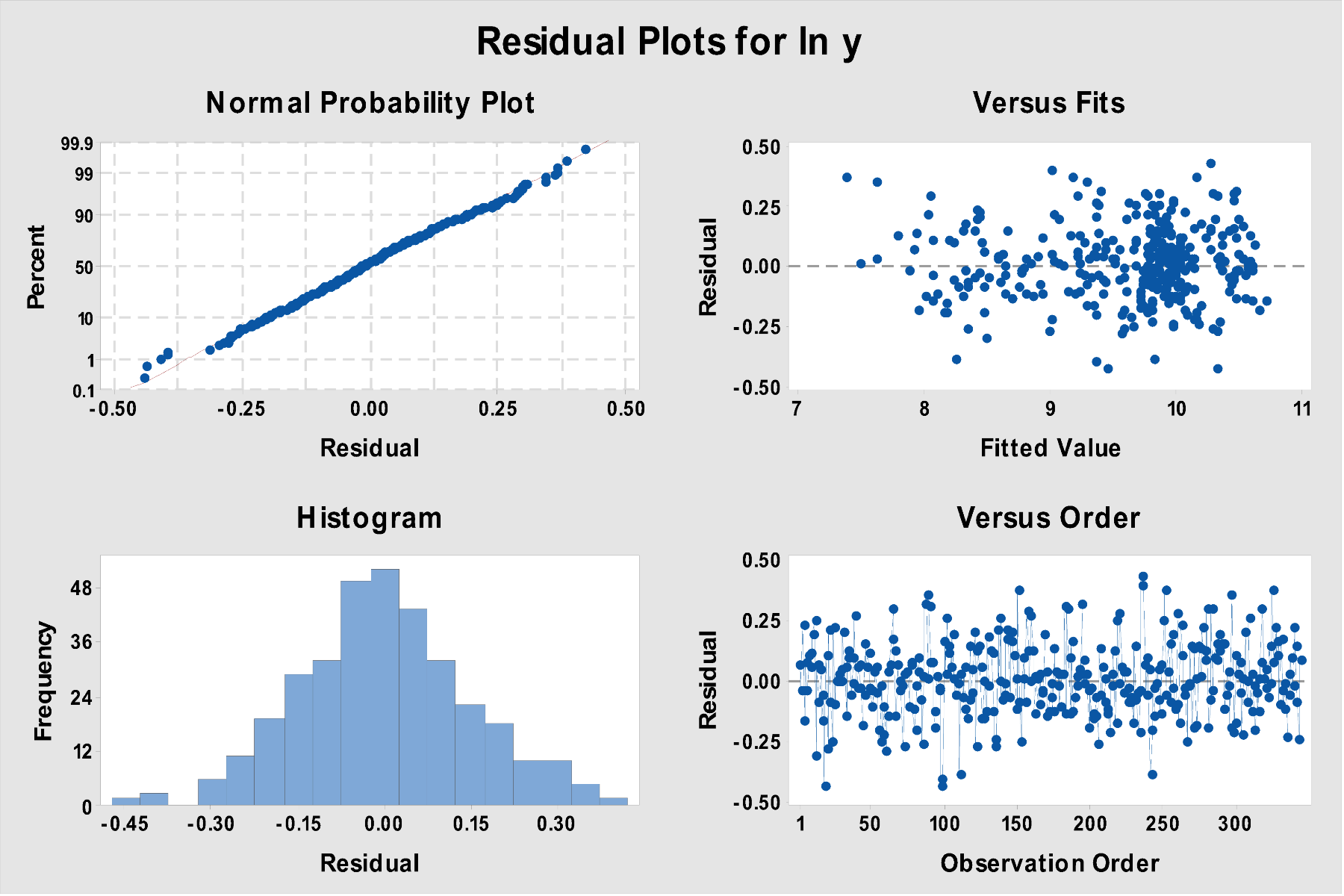


**Correlation matrix of independent variables**

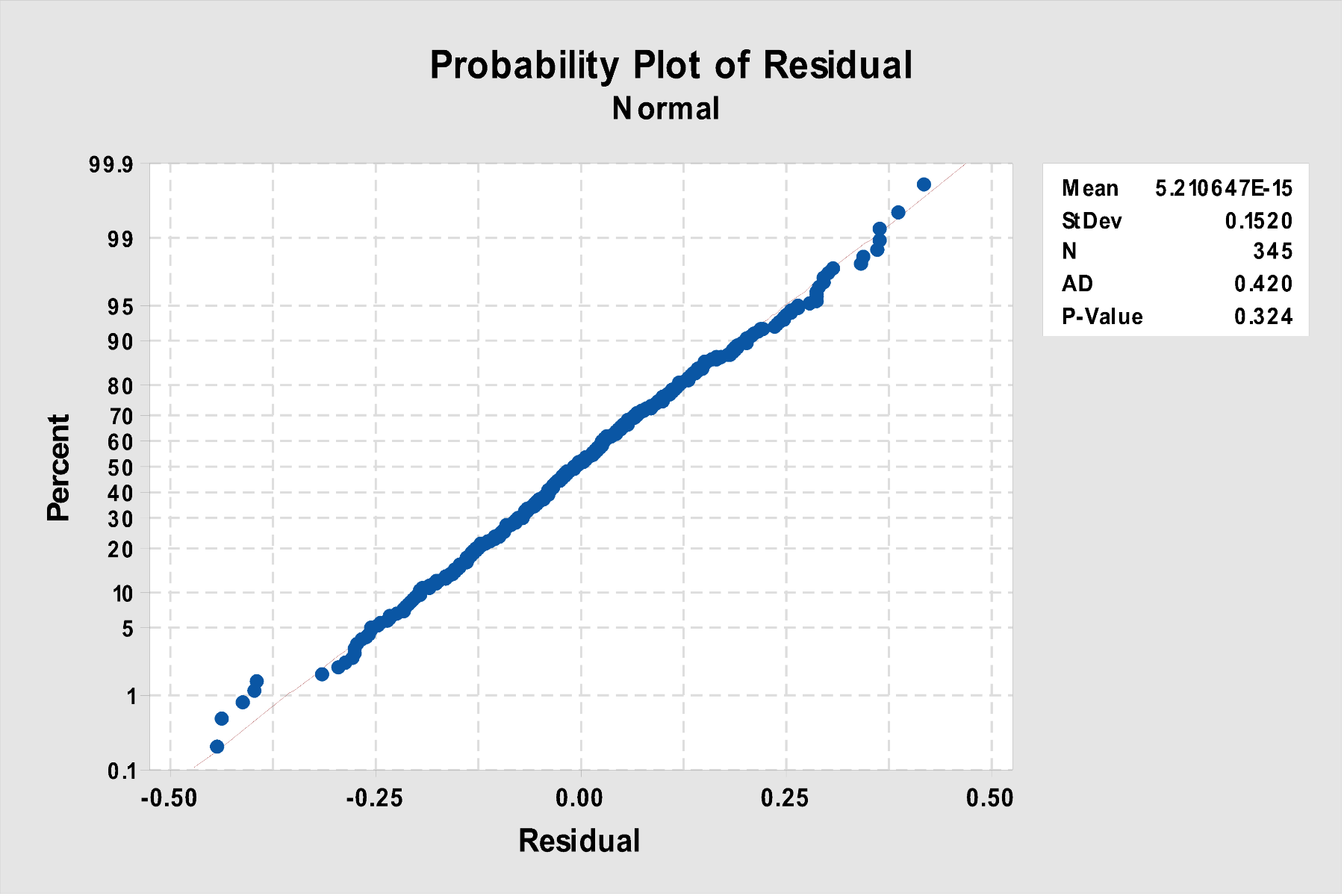
We decided to choose highly correlated variables (p-value < 0.01 in correlation matrix) and add interaction terms of them. Then we used both stepwise regression and backward elimination methods to choose the significant interaction terms in our model.

After model screening, we find that both stepwise and backward method provide the same selection of variables (the procedure of selections can be found in Appendix C1). After the second order term selection, 7 interaction variables are added to the model (*additional variables: Employment rate\*Number of rooms per person, Number of rooms per person\*Homicide rate, Number of rooms per person\*Share of households with internet, Share of labour force with at l\*Homicide rate, Share of labour force\*Voter turnout in general election, Homicide rate\*Voter turnout in general election, Homicide rate\*Share of households with internet*). Compared to the first-order model, R-squared increases from 93.96% to 95.50% and SSE reduces from 10.295 to 7.994.

Next, we checked the assumptions on residuals of the second order model.



**Residual Plots for Second Order Model**



**Normality plot with A-D Test for Second Order Model**

1. **Normality assumption**

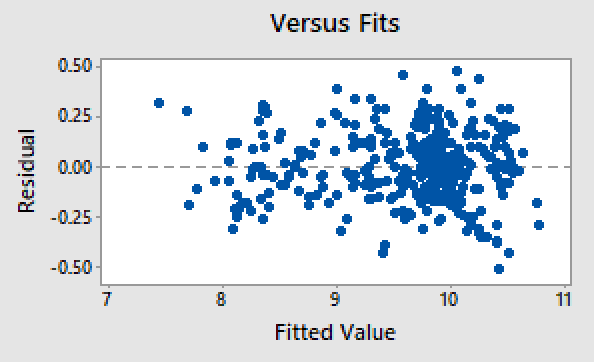
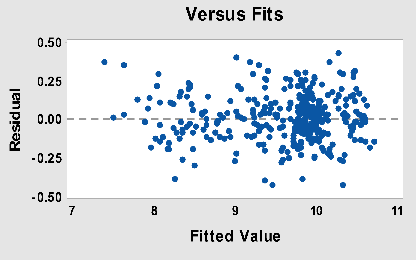
The p-value of Anderson-Darling normality test is 0.324, which is larger than 0.05, meaning that residuals are distributed normally. The histogram of residuals also provides a strong evidence that the residuals are normally distributed.

1. **Mean assumption**

The mean of residuals is 5.21064\*E-15, which is very close to zero. Therefore, we can conclude that the residual is normally distributed with mean= 0.

1. **Constant variance assumption**

In the first-order model, we find that the residual variance slightly increases as fitted value gets larger (fann-out). After adding interaction terms, the residual versus fit plot show less fan-out pattern and residuals from a horizontal band between -0.50 and 0.50, which suggests a random distribution and fulfill the constant variance assumption.

**→**

**Residual vs. Fitted Value plot Residual vs. Fitted Value plot**

**First-order model Second-order model**

1. **Independence assumption**

The Durbin-Watson statistic for the second-order model is 1.88057. The test statistics d=1.88057 > Du =1.565, which indicates that there is no autocorrelation. The residual versus order graph also does not show autocorrelation. Therefore, we can conclude that the terms are independent.

After assumption tests, we decide to perform a partial F-model test to validate the significance of the second-order model. The procedures of the test are shown below:

***Complete model: Second-order model with additional 7 interaction terms***

***Reduced model: First-order model***

Therefore, we can reject at 95% confidence level and conclude that adding additional interaction terms is significant to the model. The partial F test result suggests that the second-order model is better and should be chosen as the final model over the first-order model.

*Final Model*

Based on the the F-partial test, we decided to use the second-order model we obtained from stepwise regression considering interaction terms with highly correlated variables. This model includes 20 predictors (8 quantitative variables, 1 categorical variable (with 5 levels) and 7 interaction terms) and 345 data points. The final model is summarized below:

**Second-order Final Model**

Regression Equation:

|  |  |  |
| --- | --- | --- |
| lny | = | 5.742 - 0.01659 Employment rate + 0.479 Number of rooms per person  + 0.02334 Share of labour force with at least secondary education - 0.07035 Standardised mortality rate + 0.0381 Homicide rate + 0.02856 Voter turnout in general election + 0.02864 Share of households with internet access + 0.002509 Perception of corruption + 0.0 Continent\_Asia + 0.2435 Continent\_Europe + 0.3677 Continent\_North America + 0.0741 Continent\_Oceania + 0.3699 Continent\_South America + 0.01068 Employment rate\*Number of rooms per person + 0.02204 Number of rooms per person\*Homicide rate  - 0.00946 Number of rooms per person\*Share of households with internet access  + 0.000876 Share of labour force with at least secondary education\*Homicide rate - 0.000287 Share of labour force with at least secondary education\*Voter turnout in general election - 0.001188 Homicide rate\*Voter turnout in general election - 0.000789 Homicide rate\*Share of households with internet access |

*Model Summary*  
S R-sq R-sq(adj) R-sq(pred)  
0.156339 95.50% 95.24% 93.88%

*Coefficients*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Term | Coef | SE Coef | T-Value | P-Value | VIF |
| Constant | 5.742 | 0.442 | 12.99 | 0.000 |  |
| Employment rate | -0.01659 | 0.00394 | -4.21 | 0.000 | 22.73 |
| Number of rooms per person | 0.479 | 0.157 | 3.05 | 0.003 | 113.73 |
| Share of labour force with at l | 0.02334 | 0.00403 | 5.78 | 0.000 | 85.18 |
| Standardised mortality rate | -0.07035 | 0.00847 | -8.31 | 0.000 | 2.22 |
| Homicide rate | 0.0381 | 0.0183 | 2.08 | 0.038 | 228.01 |
| Voter turnout in general election | 0.02856 | 0.00452 | 6.32 | 0.000 | 45.70 |
| Share of households with intern | 0.02864 | 0.00281 | 10.19 | 0.000 | 40.00 |
| Perception of corruption | 0.002509 | 0.000664 | 3.78 | 0.000 | 2.23 |
| Continent |  |  |  |  |  |
| Asia | 0.000000 | 0.000000 | \* | \* | \* |
| Europe | 0.2435 | 0.0406 | 6.00 | 0.000 | 5.81 |
| North America | 0.3677 | 0.0537 | 6.84 | 0.000 | 7.51 |
| Oceania | 0.0741 | 0.0599 | 1.24 | 0.217 | 2.77 |
| South America | 0.3699 | 0.0616 | 6.00 | 0.000 | 2.79 |
| Employment rate\*Number of rooms per person | 0.01068 | 0.00239 | 4.47 | 0.000 | 187.53 |
| Number of rooms per person\*Homicide rate | 0.02204 | 0.00986 | 2.24 | 0.026 | 70.07 |
| Number of rooms per person\*Share of households with intern | -0.00946 | 0.00186 | -5.08 | 0.000 | 167.36 |
| Share of labour force with at l\*Homicide rate | 0.000876 | 0.000283 | 3.09 | 0.002 | 89.79 |
| Share of labour force with at l\*Voter turnout in general election | -0.000287 | 0.000057 | -5.01 | 0.000 | 95.24 |
| Homicide rate\*Voter turnout in general election | -0.001188 | 0.000314 | -3.78 | 0.000 | 239.06 |
| Homicide rate\*Share of households with intern | -0.000789 | 0.000197 | -4.01 | 0.000 | 35.61 |

As the final validation of our model, we performed the F-test as following:

Based on the F-statistic, we reject null hypothesis, and can conclude that the model is significant at the 95% confidence level.

*Conclusion*

In this study, we attempted to understand how environmental/non-economic factors contribute to a region’s income per capita. We began with a naive first-order model with no variable selection and transformation, which violated some regression assumptions. We then constructed a first-order model that employed variable selection and response variable transformation to improve the fitness of the model and correct model assumptions. In order to further enhance the model performance and constant variance assumption, we built a second-order model that incorporated interaction terms from first-order model. We finally obtained a model with an adjusted R-square of 95.24% with residuals that satisfy regression assumptions.

The final model is as follows:

|  |  |  |
| --- | --- | --- |
| ln y | = | 5.742 - 0.01659 Employment rate + 0.479 Number of rooms per person  + 0.02334 Share of labour force with at least secondary education - 0.07035 Standardised mortality rate + 0.0381 Homicide rate + 0.02856 Voter turnout in general election  + 0.02864 Share of households with internet access + 0.002509 Perception of corruption  + 0.0 Continent\_Asia + 0.2435 Continent\_Europe + 0.3677 Continent\_North America  + 0.0741 Continent\_Oceania + 0.3699 Continent\_South America  + 0.01068 Employment rate\*Number of rooms per person  + 0.02204 Number of rooms per person\*Homicide rate  - 0.00946 Number of rooms per person\*Share of households with internet access  + 0.000876 Share of labour force with at least secondary education\*Homicide rate  - 0.000287 Share of labour force with at least secondary education\*Voter turnout in general election - 0.001188 Homicide rate\*Voter turnout in general election  - 0.000789 Homicide rate\*Share of households with internet access |

Based on the model we developed, we identified several factors that will contribute to the economic well-being of a region: employment rate, number of rooms per person, share of labour force with at least secondary education, standardized mortality rate, homicide rate, voter turnout in general election, share of households with internet broadband access, perception of corruption. All these individual terms, except the standardized mortality rate, are positively associated with the natural logarithm of the income per capita. To better understand the real-world implication of our final model, we calculated the effect of a one unit change in independent variable on the dependent variable, as shown in table below.

For instance, when analyzing the number of rooms per person, we keep other variables constant. The number of rooms per person has coefficient of 0.479, and it has interaction terms with employment rate, homicide rate, and share of households with internet broadband access. For the sake of demonstration, we use the mean values of those variables in the dataset, 66.31, 3.41, and 69.75 respectively. Therefore, one unit increase in number of rooms per person will result in 0.479 + 0.01068 \*66.31+ 0.02204\*3.41- 0.00946 \* 69.75 = 0.6025122 unit increase in lny. To understand the effect on response variable y, we take exponential on both sides, e^0.6025122. The result shows that one unit increase in number of rooms per person will scale the value of y (income per capita) by 1.83.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | fixed value (medium) | coeff | lny | y |
| Number of rooms per person | 1.73 | 0.479 | 0.6025122 | 1.82670208 |
| internet | 69.75 | 0.02864 | 0.01224414 | 1.01231941 |
| Share of labor | 74.96 | 0.02334 | 0.00682264 | 1.00684597 |
| homicide | 3.41 | 0.0381 | 0.00612493 | 1.00614373 |
| voter | 67.96 | 0.02856 | 0.0029954 | 1.00299989 |
| corruption | 61.08 | 0.002509 | 0.002509 | 1.00251215 |
| employment rate | 66.31 | -0.01659 | 0.0018864 | 1.00188818 |
| mortality | 8.34 | -0.07035 | -0.07035 | 0.93206754 |

Based on this result, we can infer that increasing the number of rooms per person is associated with the most amount of increase in income per capita, whereas increasing the employment rate has the least effect. Reducing the standardized mortality rate can also improve the income per capita.

While our model does not investigate into the causal effects of independent variables on dependent variable, this model can serve as a reference for policy makers when making or prioritizing decisions. For example, assume there is one policy that deal with employment rate and another policy that improves housing accessibility (increasing number of rooms per person). When considering the economical consequence of such policy, the policy maker may look into this model and realize that improving the number of rooms per person are more likely to be associated with higher income per capita than improving employment rate.

If given the chance to further investigate into this topic, we would like to incorporate more social/cultural factors to differentiate regions or countries, such as the religion of a country or the level of development of a country. Implementing such variables may help us differentiate and thus taylor our model better towards varied regions. Moreover, adding those categorical variables may also mitigate the number of outliers and influential points in our model. As a result, our model accuracy and interpretation would both be improved.

1. **References**
2. Merriam-webster.com. (n.d.). *Definition of WELL-BEING*. [online] Available at: https://www.merriam-webster.com/dictionary/well-being [Accessed 26 Nov. 2018].
3. Cswe.org. (2016). Council on Social Work Education (CSWE) - Working Definition of Economic Well-Being. [online] Available at: https://www.cswe.org/Centers-Initiatives/Initiatives/Clearinghouse-for-Economic-Well-Being/Working-Definition-of-Economic-Well-Being [Accessed 26 Nov. 2018].
4. Cswe.org. (2016). Council on Social Work Education (CSWE) - Working Definition of Economic Well-Being. [online] Available at: https://www.cswe.org/Centers-Initiatives/Initiatives/Clearinghouse-for-Economic-Well-Being/Working-Definition-of-Economic-Well-Being [Accessed 26 Nov. 2018].
5. OECD (2014). OECD Statistics. [online] OECD.Stat. Available at: https://stats.oecd.org/ [Accessed 11 Nov. 2018].
6. OECD (2014). *How to Measure Regional and Local Well-being*. [online] OECDiLibrary. Available at: https://www.oecd-ilibrary.org/urban-rural-and-regional-development/how-s-life-in-your-region/how-to-measure-regional-and-local-well-being\_9789264217416-6-en [Accessed 12 Nov. 2018].
7. **Appendix**