**BST 569/869: Linear Statistical Models**

**Take-Home Final Fall 2022**

**Due: Wednesday, December 7th by 12.00pm**

**(200 total points)**

**Instructions**

* The exam consists of two parts, Parts I and II.
* Use this Word document as your template to answer all questions. Your final document should be either a Word or a PDF document – saved using the style LastName.Final.docx or LastName.Final.pdf. o For example, if I hand in my final exam, the document would be saved as Tabb.Final.pdf.
* Submit your final document via Learn.

**Honor Code**

By signing below you confirm that you have neither given nor received any unauthorized assistance on this exam. Furthermore, you agree not

to discuss this exam with anyone until the exam testing window is over. In addition, you are not to utilize online resources to answer your questions. By signing below, you also are acknowledging

that you’ve read and understand Drexel University’s Academic Integrity Policy – available [here.](https://drexel.edu/studentlife/community-standards/code-of-conduct/academic-integrity-policy)

Signature: \_\_\_\_\_\_\_\_\_\_Arpita Deb\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Print Name: \_\_\_\_\_\_\_\_\_ARPITA DEB\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Date: 12-07-2022\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Part I\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Part I\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Part I\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**Based on the article, “Individual, Interpersonal, and Institutional Level Factors Associated with the Mental Health of College Students” (Byrd and McKinney, 2011)**

In your own words, describe the: (1) motivation of the study (i.e. study purpose), (2) study aims, (3) hypotheses being tested, (4) dependent outcome of interest, (5) study population being considered, (6) independent measures of interest, (7) statistical methods utilized, (8) primary results, (9) study strengths and (10) study limitations. Consider this an extended abstract of the manuscript, where all of this information is available in the manuscript. You are NOT to include bullet points to describe all of the components above; rather, you are writing out this extended abstract with full sentences and paragraphs as needed (just like the many manuscript abstracts you have reviewed this entire quarter). This extended abstract should be no longer than 1 page, 12-point font, 1-inch margins, and single spaced. [20 points]

**Abstract:** Currently many college students have fallen prey of mental health challenges. As a result, many such students are using counselling services more frequently than before. Since the severity of symptoms have increased, treatment duration have also increased. The far-sighted goal of this study was to understand those factors that lead to mental health issues. While most of the studies have used individual factors, this study focused on individual, interpersonal and institutional level factors to assess the outcome of mental health. **Objectives:** The aim of the study was to assess the relationship of mental health with coping abilities of students, suicidal tendencies, confidence in communication, spirituality, academic self-confidence and effort, sexual orientation, and alcohol use. The study hypothesised that students would experience deteriorated mental health status as suicidal tendencies occur at an individual level, increase in responsibilities of work/life balance at interpersonal level, less interaction with faculty and presence of a racially tense campus at institutional level. Here,Mental health outcome was the dependent variable. Study population being considered were: Secondary data from the 2006 Student Development Survey. The survey was conducted at a large public research institution. The data set included 2,203 cases. **Method:** The independent variables used were: A) Individual factors: coping abilities factor, suicidal tendencies factor, confidence in communicating factor, strong spiritual identity factor, academic self-confidence factor, heterosexual orientation, and alcohol use. Heterosexual orientation was formed by combining responses from 2 items: (1) primary sexual attraction (responses included males, females, both genders, and unsure) and (2) gender (identified as male or female). B) The interpersonal level variables included intergroup awareness factor, social engagement factor, and work/life responsibilities factor. C) The institutional level variables were as follows: campus climate/tension factor, institutional satisfaction factor, and faculty interaction factor. It was suggested that the final regression resulted in a positive correlation between the variables and coping abilities; A degree of interrelatedness (multicollinearity) among them was also noticed. The statistical methods utilized: Stepwise multiple linear regression- to estimate the impact of individual, interpersonal, and institutional level factors on overall mental health.

**Primary results**: After accounting for demographic and background characteristics, it was noted that individual and institutional level factors explained 49% (adjusted *R*2 = .48) of the variation in mental health. GPA was the only background characteristics that was significantly correlated with mental health. For the final regression model, English as your first language was selected from the demographic and background characteristics. coping abilities among the domain factors had the largest influence on mental health. When added to the model it increased R square. Overall, there was a significant change in *R*2 after entry of each individual and institutional level variable (see Table 3). The strength of the study was- it utilized step-wise regression for analysis. This allowed the researcher to manage large amounts of potential predictor variables, fine-tuning the model to choose the best predictor variables from the available options. Watching the order in which variables are removed or added can provide valuable information about the quality of the predictor variables. This study focused on investigating the individual, interpersonal, and institutional level factors associated with overall mental health among college students whereas majority of studies focus/investigate the individual factors to understand its association with overall mental health among college students.

**Study limitations:** One can detect several limitations in this study. First, the sample collection was from a public university located in western United States. Sampling lacks generalizability as they did not include private institutions and community colleges nor did they consider eastern or central Unites States. Second, There was unsatisfactory survey response (7.1%) for data collection. This may lead to potential sampling bias. Third, In this study, cross-sectional data was used that refrains us from cause-effect analysis rather longitudinal data could have been used that could evaluate causality and assert temporal change in mental health. Fourth, Due to the use of secondary data there could be construct validity. Fifth, the questionnaire did not include sexual identity. Thus, transsexual and transgender individuals were not included in the study. Sixth, Gender specification was not labelled for 1- and 5-drink measures of alcohol use. Seventh, this study demonstrates same binge drinking cut points for both men and women. It will mislead the results because it underestimates the number of female binge drinkers.

# \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Part II\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Part II\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Part II\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

The County Health Rankings & Roadmaps (CHR&R) is a program of the University of Wisconsin Population Health Institute. The CHR&R program provides data, evidence, guidance, and examples to build awareness of the multiple factors that influence health and support leaders in growing community power to improve health equity. The data is provided on a yearly basis for all counties in the US, and is publicly available for researchers, policy makers, and community leaders to understand how communities are performing in terms of health outcomes and health factors.

More information is available here: [https://www.countyhealthrankings.org.](https://www.countyhealthrankings.org/)

Your goal is to conduct an analysis that aims to understand the associations between health factors and health outcomes according to the data you have available provided by the CHR&R – for the state of Pennsylvania (N = 65 counties). Your outcome of interest is length of life, which is measured by a variable called “Years of Potential Life Lost Rate”. The potential health factors can include measures from the following four domains: health behaviors, clinical care, social and economic factors, and physical environment. The CHR&R has a conceptual model that pulls together the relationship between health outcomes and health factors – displayed below.

While the CHR&R’s conceptual model includes all four domains of health factors to predict health outcomes, your analysis will only focus on **one** domain of health factors. Therefore, you can choose to pick measures from either health behaviors, clinical care, social and economic factors, or physical environment. See this [website](https://www.countyhealthrankings.org/explore-health-rankings/county-health-rankings-model) for an interactive display that shows the various measures related to each of the four domains. Additionally, demographic measures like age (% below 18 years of age, % 65 and older), race/ethnicity (i.e. % African American, % American Indian and Alaska Native, % Asian, % Hispanic, % non-Hispanic white), % not proficient in English, % females, and % rural are also provided.

Diagram

Description automatically generated

**The data and codebook are provided in the file “CHR2022PA.xlsx”, and a tab delimited file also includes the data, called “CHR2022PA.txt”. To import the data easily, you can access the saved workspace using the file: “CHR2022PA.RData”.**

Your final report will include three parts: o Sections 1 and 2: MAIN REPORT – This section should include all of the sections and subsections described below and should be no longer than 3 pages, 12-point font, 1-inch margins, and single spaced – there should be NO R code/output in this section. o Section 3: TABLES/FIGURES - All tables/figures need to be labeled as “Table XX.

Descriptive table name”; “Figure XX. Descriptive figure name” – there should be NO R code/output in this section. There are no page limits for the tables/figures; however, you should take care to create figure that are not too large, but not too small as to lose clarity and interpretability.

o Section 4: APPENDIX- This section should include all R code associated with your analyses. The font in this section can be an 8-point font to save space - there should be NO R output in this section.

* Round all numbers to the 3rd decimal place.
* Label all sections in your report according to the Sections provided below.
* If you do not adhere to these instructions, you will lose points (no exceptions).

**Section 1: METHODS**

# Sub-Section 1.1: Descriptive statistics [10 points]

- This section should describe the descriptive statistics you will examine, as it applies to premature death (which is the main outcome of interest) as well as one of the four domains, and the demographic measures.

# Sub-Section 1.2: Simple linear regression [15 points]

- This section should describe the univariate analyses (i.e. marginal relationships) that you will examine between premature mortality and the various health factors, as well as premature mortality and the various demographic measures.

# Sub-Section 1.3: Multiple linear regression [10 points]

- This section should describe the proposed multivariable analyses (i.e. joint relationships) that you will examine between premature mortality and the various health factors and demographics.

# Sub-Section 1.4: Interaction models [10 points]

- This section should describe the proposed interaction model, that considers different associations based on at least one demographic factor. Because there are a vast number of possible interaction effects, your job here is to only build one interaction model.

# Sub-Section 1.5: ANOVA [10 points]

- This section should describe an analysis of variance – taking into consideration the models examined in the previous Sub-Sections.

# Sub-Section 1.6: Transformations [10 points]

- This section should describe your considerations for possible transformations for your outcome of interest, as well as your independent measures of interest.

# Sub-Section 1.7: Regression Diagnostics [15 points]

- This section should describe residual analyses, non-constant variance, outliers/influential points, as well as the normality assumption of your final model.

# Section 2: RESULTS [80 points]

- This entire section should present your results and interpretations of your analyses based on the Methods section described above. All of the text in this Results section should follow logically based on the Methods details provided in Section 1.

# Section 3: TABLES AND FIGURES [10 points]

# Section 4: APPENDIX – R CODE (No Output!!) [10 points]

**Section 1: METHODS**

# Sub-Section 1.1: Descriptive statistics [10 points]

-Descriptive statistics: There are 65 observations and 49 variables.

- In this census population estimates in the year 2020, The Health Factor- **Physical Environment** is my area of interest. Physical environment has two sections:

**1. Air & Water Quality**. Subdivided into a. Air Pollution- Particulate Matter= (numeric) continuous variable, b. Drinking water violations= categorical variable.

**2. Housing & Trait**. Subdivided into a. Severe Housing Problems = (integer) discrete variable. b. Driving Alone to Work= (integer) discrete variable. c. Long Commute-Drive Alone = (integer) discrete variable. The outcome= Premature mortality= (integer) discrete variable.

The demographic variables: Population (integer- discrete variable), %Below 18 Years of age(numeric), % 65 and Older (numeric), % Non-Hispanic Black( numeric), % American Indian & Alaska Native(numeric), % American Indian & Alaska Native(numeric), % Asian(numeric), % Native Hawaiian/Other Pacific Islander(numeric), % Hispanic(numeric), % Non-Hispanic White(numeric), % Not Proficient in English, % Female(integer- discrete variable), % rural(numeric). All numeric variables are continuous variables. To evaluate the descriptive statistics, a summary of mean and standard deviation was found for the physical environment(domain) variables and demographic variables, as presented in Table 1 and 2.

As a part of descriptive statistics analysis, a correlation matrix was also developed, presented in table 3and 4. In addition a scatterplot matrix was also plotted (graph 1 and 2) to visualize the correlation of the variables with themselves and with the outcome- premature mortality.

# Sub-Section 1.2: Simple linear regression [15 points]

Simple linear regressions were carried out with each of the independent variable- domain: Physical Environment and demographic measures against the outcome (dependent variable): premature mortality, as depicted in table 5 and 6. Coefficient, standard error, T-test and p-value of each independent variable was evaluated with the outcome variable- Premature mortality, keeping the p-value level of significance at 0.05.

# Sub-Section 1.3: Multiple linear regression [10 points]

Next, two joint relationship/ multivariate analysis was carried out: First, amongst Physical Environment (independent variable) and premature mortality (dependent variable) as shown in Table 8. Second, between demographic variables (independent variable) and premature mortality (dependent variable) as shown in Table 7. Coefficient, standard error, T-test and p-value of each independent variable was evaluated with the outcome variable- Premature mortality, keeping the p-value level of significance at 0.05.

# Sub-Section 1.4: Interaction models [10 points]

As a next step, I aimed to evaluate the interaction between the variables. Since the domain variables and demographic variables are large in count, so to select the variables for interaction model the following steps were undertaken:

1. Reviewed the correlation matrix
2. Reviewed the outcome of multiple linear regression model
3. Considered a Forward stepwise model where the AIC value was considered (Table 18 and 19)

After evaluating all the three measures explained above, I went ahead and considered my final model with the two variables: Long Commute-Drive Alone and **%** Below 18 Years of age, presented in Table 9. Their interaction effect on premature mortality (dependent variable) was evaluated as presented in Table 10.

# Sub-Section 1.5: ANOVA [10 points]

Anova type-II was carried out to analyse the most complex model to the least complex model, and finally the main effect model. The Table 11 depicts the result of Anova type II of all the models along with the interaction model. Table 12. depicts the result of Anova type II of all the models with the final variables incorporated in it. Other Anova models with the main effects were also evaluated (as shown in R codes in Appendix). They are not highlighted in the table.

# Sub-Section 1.6: Transformations [10 points]

Next a transformation for outcome premature mortality, as well as Long Commute-Drive Alone (Table 4) and **%** Below 18 Years of age variables measures (Table 5) were carried out.

# Sub-Section 1.7: Regression Diagnostics [15 points]

As a part of regression analysis, regression plots were graphed and evaluated (Table 15) and Figure 6. Nonconstant variance (Table 16), outliers (Table 17) and graph, normality assumption plot- QQ plot in figure 7.

# Section 2: RESULTS [80 points]

Descriptive statistics that includes mean and standard deviation of outcome- Premature mortality, domain variables from Physical environment for the study participants are as follows: The mean of Population was 386227.2 and standard deviation was 1585824. Mean of population % Below 18 Years of age was 19.672 and standard deviation was 2.027. Mean of population % 65 and Older was 21.149 and standard deviation was 2.511 and mean of % Non-Hispanic Black was 4.462 and standard deviation was 6.054 the mean % American Indian & Alaska Native is 0.332 and standard deviation was 0.190 the mean of % Asian was 1.795 and standard deviation was 2.040; the mean % Native Hawaiian/Other Pacific Islander was 0.055 and standard deviation was 0.061; the mean % Hispanic was 4.985 and standard deviation was 5.518; the mean of % Non-Hispanic White was 87.194 and standard deviation was 11.389. The mean of % Not Proficient in English between the year 2016-2020 was 0.923 and standard deviation was 1.254; mean of % Female was 50.186 and standard deviation was 1.310; Similarly, the mean of % Rural was 46.575 and the standard deviation was 27.218, also presented in Table 1. and Descriptive statistics of demographic and background characteristics are presented in Table 2. Next, to understand the correlation between the variables better a correlation matrix, presented in table 3 and 4, and scatterplot matrix, (presented in Figure 1 and 2) was evaluated. From the scatterplot graph and correlation table we find that none of the variables have a strong correlationwith one another or with outcome-premature mortality.

Next, to analyse relationship between each independent variable with outcome-premature mortality, simple linear regression analysis was carried out, as seen in table 5 and 6. The results of each regression analysis was found to be not statistically significant (p<0.05). That means none of the Physical Environment and demographic variables were statistically significant. Moreover, the R2 value was low for all the variables suggesting that not much of the variation of the dependent variable is explained by the independent variables in the simple linear regression.

To analyse the data further, multiple linear regression was carried out. First, amongst Physical Environment (independent variable) and premature mortality (dependent variable) as shown in Table 8. Second, between demographic variables (independent variable) and premature mortality (dependent variable) as shown in Table 7. It was found that none of the variables were statistically significant when p-value (level of

significance) was kept at <0.05, and R2 value was low in both the models (Physical environment and demographic) suggesting that not much of the variation of the dependent variable is explained by the independent variables in the simple linear regression.

To authenticate the model further, I aimed to evaluate the interaction between the variables. To select the final model initially p-value of all the independent variables were considered. Since the p-value of the variables were insignificant, selection of variables were inconclusive. Similarly, when the correlation table was referred, I could not draw any conclusion on the variable selection. Finally, a forward stepwise regression method was used. Based on the AIC value a final model was generated with Long Commute-Drive Alone and **%** Below 18 Years of age, and their interaction effect on premature mortality (outcome) was evaluated. (Table 9). Based on this method, the variables from physical environment domain variables- Air Pollution- Particulate Matter, Severe Housing Problems, Driving Alone to Work and Long Commute Driving Alone were eliminated.

The regression summary of the interaction model (Table 10) showed that the interaction model was statistically insignificant, considering p-value(<0.05). Infact, the main effect model of Long Commute-Drive Alone on premature mortality and **%** Below 18 Years of age on premature mortality was also found to be insignificant, considering p-value(<0.05). This suggests that there was no interaction present between the variables **%** Below 18 Years of age and Driving Alone to Work. So we do not need to complicate the model any further.

In the following step, after fitting the mean functions, based on the marginality principle: Anova (Type II test) began with highest order of interaction to the lowest. I evaluated the Anova tables starting from most complex to least complex. However, the result of Anova model (Table 11) model showed that the values were found to be insignificant, considering p-value(<0.05). Since, the p-value is not significant we fail to reject the null hypothesis. Similarly, with the model depicted in Table 12, even with the main effect model. All had p value >0.05. Hence we can conclude that the hypothesised Anova models (presented in appendix) cannot explain the impact of the predictors on the response variable respectively.

Transformation model that was formed to fit the model better as seen in Table 13 and Graph 5 between **%** Below 18 Years of age on premature mortality. Table 13 is a table of RSS as a function of the transformation parameter λ, evaluated at λ ∈ (−1, 0, 1) and at the value of λ that minimizes RSS, obtained by solving a nonlinear regression problem. In the example λˆ = 0.05 ≈ 0, so the log-transform for **%** Below 18 Years of age is suggested. Next, transformation was done between Long Commute-Drive Alone on premature mortality (Table 14, Graph 4.) Table 14 is a table of RSS as a function of the transformation parameter λ, evaluated at λ ∈ (−1, 0, 1) and at the value of λ that minimizes RSS, obtained by solving a nonlinear regression problem. In the example λˆ = 0.05 ≈ 0, so the log-transform for Long Commute-Drive Alone is suggested.

The residual analysis using the regression plot in Figure 3, Table 15, 16, 17, showed the trends in our residuals, evidence of heteroskedasticity and possible outliers. The plot for this model indicates that our model is roughly systematically with skewness towards negative direction in the right of the curve. Although the residuals do not seem to be evenly spread around 0 for all fitted values, the range of the residuals at each fitted value appears to be roughly the same. However, there is evidence of heteroskedasticity in some spaces. Finally, this plot indicates that there is presence of outliers because there are 3 points on the plot that are well-separated from the rest. In addition, the residual plots in figure 6 shows that in the graph with Long Commute-Drive Alone vs premature mortality, the samples are in a randomly equally scattered above and below central line. This random pattern indicates that a linear model provides a decent fit to the data. Whereas for the variable- between **%** Below 18 Years of age and the fitted model, the samples are skewed to the right.

The next plot is the QQ-plot (Figure7). Though most of the points seem to fall on the line which indicates that our residuals come from a normal distribution, there are some points that stray from the line in the lower quantiles of the plot. There is also evidence of outliers.

# Section 3: TABLES AND FIGURES [10 points]

**Table 1- Descriptive characteristics of Physical Environment in health factor**

|  |  |  |
| --- | --- | --- |
| **Table 1.** | **Descriptive characteristics of variable:** | **Physical Environment in health factor** |
|  | Mean | Standard Deviation |
| **Air Pollution- Particulate Matter** | 8.411 | 1.193 |
| **Severe Housing Problems** | 12.51 | 2.575 |
| **Driving Alone to Work** | 79.2 | 4.887 |
| **Long Commute Driving Alone** | 35.31 | 10.082 |
| **Drinking Water Violation** | Frequency | Proportion |
| **Yes** | 52 | 0.8 |
| **No** | 13 | 0.2 |

**Table 2- Descriptive characteristics Demographic measures**

|  |  |  |
| --- | --- | --- |
| **Table 2.** | **Descriptive characteristics of variable** | **Demographic measures** |
|  | Mean | Standard Deviation |
| **Population** | 386227.2 | 1585824 |
| **% Below 18 Years of age** | 19.672 | 2.027 |
| **% 65 and Older** | 21.149 | 2.511 |
| **% Non-Hispanic Black** | 4.462 | 6.054 |
| **% American Indian & Alaska Native** | 0.332 | 0.190 |
| **% Asian** | 1.795 | 2.040 |
| **% Native Hawaiian/Other Pacific Islander** | 0.055 | 0.061 |
| **% Hispanic** | 4.985 | 5.518 |
| **% Non-Hispanic White** | 87.194 | 11.389 |
| **% Not Proficient in English** | 0.923 | 1.254 |
| **% Female** | 50.186 | 1.310 |
| **% Rural** | 46.575 | 27.218 |
|  |  |  |

**Table 3- Correlation matrix Part I**

|  |  |  |
| --- | --- | --- |
| **Table 3** | **Premature mortality** | **Population** |
| **Premature mortality** | **1.0** | -0.110 |
| **Population** | **-0.110** | 1.0 |
| **% Below 18 Years of age** | **0.157** | 0.099 |
| **% 65 and Older** | **-0.036** | -0.198 |
| **% Non-Hispanic Black** | **0.010** | 0.266 |
| **% American Indian & Alaska Native** | **0.044** | 0.102 |
| **% Asian** | **-0.067** | 0.258 |
| **% Native Hawaiian/Other Pacific Islander** | **0.030** | 0.162 |
| **% Hispanic** | **0.014** | 0.139 |
| **% Non-Hispanic White** | -0.003 | -0.261 |
| **% Not Proficient in English** | **0.053** | 0.219 |
| **% Female** | **0.031** | 0.154 |
| **% Rural** | **-0.019** | -0.232 |

**Table 4- Correlation matrix Part II**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 4** | **% Below 18 Years of age** | **% 65 and Older** | **% Non-Hispanic Black** | **% American Indian & Alaska Native** | **% Asian** | **% Native Hawaiian/Other Pacific Islander** | **% Hispanic** | **% Non-Hispanic White** | **% Not Proficient in English** | **% Female** | **% Rural** |
| **premature mortality** | 0.157 | -0.037 | 0.010 | 0.043 | -0.067 | 0.030 | 0.0137 | -0.003 | 0.053 | 0.031 | -0.019 |
| **Population** | 0.099 | -0.198 | 0.267 | 0.102 | 0.258 | 0.161 | 0.139 | -0.261 | 0.219 | 0.154 | -0.232 |
| **% Below 18 Years of age** | 1.0 | -0.532 | 0.183 | 0.195 | 0.246 | 0.327 | 0.343 | -0.316 | 0.376 | 0.554 | -0.394 |
| **% 65 and Older** | -0.532 | 1.0 | -0.551 | -0.341 | -0.694 | -0.588 | -0.499 | 0.675 | -0.600 | -0.255 | 0.661 |
| **% Non-Hispanic Black** | 0.183 | -0.551 | 1.0 | 0.411 | 0.687 | 0.403 | 0.400 | -0.871 | 0.640 | 0.319 | -0.542 |
| **% American Indian & Alaska Native** | 0.195 | -0.341 | 0.411 | 1.0 | 0.262 | 0.540 | 0.873 | -0.693 | 0.776 | 0.032 | -0.219 |
| **% Asian** | 0.246 | -0.694 | 0.687 | 0.262 | 1.0 | 0.468 | 0.414 | -0.767 | 0.633 | 0.322 | -0.686 |
| **% Native Hawaiian/Other Pacific Islander** | 0.327 | -0.588 | 0.403 | 0.540 | 0.468 | 1.0 | 0.640 | -0.620 | 0.646 | 0.226 | -0.516 |
| **% Hispanic** | 0.343 | -0.499 | 0.400 | 0.873 | 0.414 | 0.640 | 1.0 | -0.780 | 0.827 | 0.205 | -0.488 |
| **% Non-Hispanic White** | -0.316 | 0.675 | -0.871 | -0.693 | -0.767 | -0.620 | -0.780 | 1.000 | -0.865 | -0.345 | 0.671 |
| **% Not Proficient in English** | 0.376 | -0.600 | 0.640 | 0.776 | 0.633 | 0.646 | 0.828 | -0.865 | 1.000 | 0.276 | -0.573 |
| **% Female** | 0.554 | -0.255 | 0.320 | 0.032 | 0.322 | 0.226 | 0.205 | 0.205 | 0.276 | 1.000 | -0.492 |
| **% Rural** | -0.395 | 0.661 | 0.542 | -0.219 | -0.686 | -0.516 | -0.488 | -0.488 | -0.573 | -0.492 | 1.000 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 5** | **Estimate** | **Std. Error** | **t value** | **Pr(>|t|)** | **R-squared** | **Adjusted R-squared** | **F-statistic** |  |
| **Air Pollution- Particulate Matter** | -44.34 | 147.20 | -0.301 | 0.764 | 0.001 | -0.014 | 0.091 |  |
| **Severe Housing Problems** | -6.299 | 68.248 | -0.092 | 0.927 | 0.000 | -0.016 | 0.009 |  |
| **Driving Alone to Work** | 23.48 | 35.84 | 0.655 | 0.515 | 0.007 | -0.009 | 0.429 |  |
| **Long Commute Driving Alone** | 12.41 | 17.36 | 0.715 | 0.477 | 0.008 | -0.008 | 0.511 |  |
| **Drinking Water Violation (Yes)** | 79.69 | 435.81 | 0.183 | 0.855 | 0.001 | -0.015 | 0.033 |  |
|  |  |  |  |  |  |  |  |  |

**Table 5-Simple Linear Regression with Domain- Physical environment**

**Table 6-Simple Linear Regression with Demographic Measures**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 6** | **Estimate** | **Std. Error** | **t value** | **Pr(>|t|)** | **R-squared** | **Adjusted R-squared** | **F-statistic** |  |
| **Population** | -9.663E-05 | 1.101E-04 | -0.877 | 0.384 | 0.012 | -0.004 | 0.770 |  |
| **% Below 18 Years of age** | 107.9 | 85.6 | 1.260 | 0.212 | 0.025 | 0.009 | 1.588 |  |
| **% 65 and Older** | -20.04 | 69.92 | -0.287 | 0.775 | 0.001 | -0.015 | 0.082 |  |
| **% Non-Hispanic Black** | 2.27 | 29.03 | 0.078 | 0.938 | 9.703E-05 | -0.016 | 0.006 |  |
| **% American Indian & Alaska Native** | 321.9 | 921.6 | 0.349 | 0.728 | 0.002 | -0.014 | 0.122 |  |
| **% Asian** | -45.22 | 85.97 | -0.526 | 0.601 | 0.004 | -0.011 | 0.277 |  |
| **% Native Hawaiian/Other Pacific Islander** | 687.5 | 2864.6 | 0.24 | 0.811 | 0.001 | -0.015 | 0.058 |  |
| **% Hispanic** | 3.472 | 31.844 | 0.109 | 0.914 | 0.000 | -0.016 | 0.012 |  |
| **% Non-Hispanic White** | -0.384 | 15.429 | -0.025 | 0.98 | 9.829E-06 | -0.016 | 0.001 |  |
| **% Not Proficient in English** | 59.2 | 139.9 | 0.423 | 0.674 | 0.003 | -0.013 | 0.179 |  |
| **% Female** | 32.55 | 134.05 | 0.243 | 0.809 | 0.001 | -0.015 | 0.059 |  |
| % Rural | -0.968 | 6.455 | -0.15 | 0.881 | 0.000 | -0.016 | 0.022 |  |

**Table 7- Multiple Linear Regression with Demographic Measures**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 7** | **Estimate** | **Std. Error** | **t-value** | **Pr(>|t|)** |
| **(Intercept)** | 9.064E+04 | 6.897E+04 | 1.314 | 0.195 |
| **Population** | -9.677E-05 | 1.216E-04 | -0.796 | 0.430 |
| **% Below 18 Years of age** | 1.277E+02 | 1.421E+02 | 0.899 | 0.373 |
| **% 65 and Older** | -1.837E+01 | 1.437E+02 | -0.128 | 0.899 |
| **% Non-Hispanic Black** | -8.146E+02 | 6.922E+02 | -1.177 | 0.245 |
| **% American Indian & Alaska Native** | 3.42E+01 | 2.879E+03 | 0.012 | 0.991 |
| **% Asian** | -1.012E+03 | 7.433E+02 | -1.362 | 0.179 |
| **% Native Hawaiian/Other Pacific Islander** | -3.407E+01 | 4.366E+03 | -0.008 | 0.994 |
| **% Hispanic** | -8.681E+02 | 6.698E+02 | -1.296 | 0.201 |
| **% Non-Hispanic White** | -7.978E+02 | 6.674E+02 | -1.195 | 0.237 |
| **% Not Proficient in English** | 4.782E+02 | 4.316E+02 | 1.108 | 0.273 |
| **% Female** | -1.242E+02 | 2.079E+02 | -0.597 | 0.553 |
| **% Rural** | 3.956E-01 | 1.381E+01 | 0.029 | 0.977 |
| **R-squared** | 0.098 |  |  |  |
| **Adjusted R-squared** | -0.110 |  |  |  |
| **F-statistic** | 0.471 |  |  |  |

**Table 8- Multiple Linear Regression with Physical environment**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 8** | **Estimate** | **Std. Error** | **t- value** | **Pr(>|t|)** |
| **(Intercept)** | 3576.598 | 5461.9746 | 0.655 | 0.515 |
| **Air Pollution- Particulate Matter** | -16.050 | 167.5184 | -0.096 | 0.924 |
| **Severe Housing Problems** | 35.177 | 95.0886 | 0.370 | 0.713 |
| **Driving Alone to Work** | 40.512 | 53.1011 | 0.763 | 0.449 |
| **Long Commute Driving Alone** | 15.066 | 18.286 | 0.824 | 0.413 |
| **Drinking Water Violation (Yes)** | 0.544 | 475.046 | 0.001 | 0.999 |
| **R-squared** | 0.021 |  |  |  |
| **Adjusted R-squared** | -0.062 |  |  |  |
| **F-statistic** | 0.256 |  |  |  |

**Table 9-Final model for regression analysis**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 9** | **Estimate** | **Std. Error** | **T value** | **Pr(>|t|)** |
| **(Intercept)** | 4866.77 | 1859.61  \* | 2.617 | 0.011\* |
| **Driving Alone to Work** | 14.43 | 17.32 | 0.833 | 0.408 |
| **% Below 18 Years of age** | 114.18 | 86.15 | 1.325 | 0.190 |
| **R-squared** | 0.035 |  |  |  |
| **Adjusted R-squared** | 0.004 |  |  |  |

\*p value significant (<0.05)

**Table 10-Interaction model**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 10** | **Estimate** | **Std. Error** | **t-value** | **Pr(>|t|)** |
| **(Intercept)** | -1799.598 | 8350.028 | -0.216 | 0.830 |
| **Driving Alone to Work** | 186.516 | 210.819 | 0.885 | 0.380 |
| **% Below 18 Years of age** | 442.881 | 410.512 | 1.079 | 0.285 |
| **Driving Alone to Work : % Below 18 Years of age** | -8.481 | 10.355 | -0.819 | 0.416 |
| **R-squared** | 0.046 |  |  |  |
| **Adjusted R-squared** | -0.001 |  |  |  |

**Table 11- Anova Type-II with the interaction model**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 11 Anova (Type II tests)** | **Sum Sq** | **Df** | **F-value** | **Pr(>F)** |
| **Driving Alone to Work** | 1344609 | 1 | 0.694 | 0.408 |
| **% Below 18 Years of age** | 3403155 | 1 | 1.757 | 0.190 |
| **Driving Alone to Work : % Below 18 Years of age** | 1306454 | 1 | 0.674 | 0.415 |
| **Residuals** | 120104094 | 62 |  |  |

**Table 12- Anova Type-II with the dependent variables -model**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 12- Anova (Type II tests)** | Sum Sq | Df | F value | Pr(>F) |
| **Driving Alone to Work** | 1344609 | 1 | 0.694 | 0.408 |
| **% Below 18 Years of age** | 3403155 | 1 | 1.757 | 0.190 |
| **Residuals** | 120104094 | 62 |  |  |

**Table 13- Transformation with Premature mortality and % Below 18 Years of age**

|  |  |  |
| --- | --- | --- |
| **Table 13** | **lambda** | **RSS** |
| **1** | 9.910 | 121024176 |
| **2** | -1.000 | 122215489 |
| **3** | 0.000 | 121756340 |
| **4** | 1.000 | 121448703 |

**Table 14- Transformation with Premature mortality and Driving Alone to Work**

|  |  |  |
| --- | --- | --- |
| **Table 14** | **lambda** | **RSS** |
| **1** | 4.354 | 123105886 |
| **2** | -1.000 | 123928594 |
| **3** | 0.000 | 123719230 |
| **4** | 1.000 | 123507249 |

**Table 15- Summary of Residual Plots**

|  |  |  |
| --- | --- | --- |
| **Table 15** | **Test stat** | **Pr(>|Test stat|)** |
| **Driving Alone to Work** | 0.117 | 0.908 |
| **% Below 18 Years of age** | 0.258 | 0.797 |
| **Tukey test** | -0.076 | 0.939 |

**Table 16- Non-constant Variance**

|  |  |  |  |
| --- | --- | --- | --- |
| **Non-constant Variance** | **Score** | **Test** |  |
| **Variance formula:** | ~ fitted.values |  |  |
| **Chisquare:** | 0.069 | Df = 1, | p = 0.792 |

**Table 17- Summary of Outliers**

|  |  |
| --- | --- |
| **No Studentized residuals with Bonferroni p < 0.05** |  |
| **Largest |rstudent|:** |  |
| **rstudent unadjusted p-value Bonferroni p** |  |
| **13 -2.502847           0.015.   0.976** |  |
|  |  |

**Table 18- Forward Stepwise Regression of Physical Environment**

A picture containing chart

Description automatically generated

A picture containing table

Description automatically generated

**Table 19- Forward Stepwise Regression of Demographic Measures**

Table

Description automatically generated

**Figure 1 - Scatterplot of domain- Physical environment with Outcome: Premature mortality**

Graphical user interface

Description automatically generated with low confidence

**Figure 2- Scatterplot of Demographic variables with Outcome: Premature mortality**

Table

Description automatically generated

**Figure 3- Residual graph of Final fitted model: Premature mortality with Driving Alone to Work and % Below 18 Years of age**

Chart, scatter chart

Description automatically generated

**Figure 4- Transformation graph of outcome Premature mortality with Driving Alone to Work**

Chart, scatter chart

Description automatically generated

**Figure 5- Transformation graph of outcome Premature mortality with % Below 18 Years of age**

Chart, scatter chart

Description automatically generated

**Figure 6- residual plot of final model**

Diagram, engineering drawing, schematic

Description automatically generated

**Figure 7- Normality Assumption q-q Plot**

Chart, scatter chart

Description automatically generated

# Section 4: APPENDIX – R CODE (No Output!!) [10 points]

install.packages("alr4")

install.packages("car")

library(car)

library(alr4)

library(fbasics)

library(tidyverse)

library(caret)

library(leaps)

library(olsrr)

load("/Users/arpitadeb/Library/CloudStorage/OneDrive-DrexelUniversity/Biostatistics Books/569-linear-regression contents/FINALS/CHR2022PA.RData")

data(CHR2022PA)

attach(CHR2022PA)

head(CHR2022PA)

str(CHR2022PA)

#summary of the dataset

summary(CHR2022PA)

summary(X..Severe.Housing.Problems)

#standard deviation calculation

sd(X..Severe.Housing.Problems)

sd(Average.Daily.PM2.5)

sd(X..Drive.Alone.to.Work)

sd(X..Long.Commute...Drives.Alone)

#Descriptive statistics of demographic measures: mean & standard deviation

mean(CHR2022PA$Population)

sd(CHR2022PA$Population)

mean(CHR2022PA$X..Less.Than.18.Years.of.Age)

sd(CHR2022PA$X..Less.Than.18.Years.of.Age)

mean(CHR2022PA$X..65.and.Over)

sd(CHR2022PA$X..65.and.Over)

mean(CHR2022PA$X..Black)

sd(CHR2022PA$X..Black)

mean(CHR2022PA$X..American.Indian...Alaska.Native)

sd(CHR2022PA$X..American.Indian...Alaska.Native)

mean(CHR2022PA$X..Asian)

sd(CHR2022PA$X..Asian)

mean(CHR2022PA$X..Native.Hawaiian.Other.Pacific.Islander)

sd(CHR2022PA$X..Native.Hawaiian.Other.Pacific.Islander)

mean(CHR2022PA$X..Hispanic)

sd(CHR2022PA$X..Hispanic)

mean(CHR2022PA$X..Non.Hispanic.white)

sd(CHR2022PA$X..Non.Hispanic.white)

mean(CHR2022PA$X..Not.Proficient.in.English)

sd(CHR2022PA$X..Not.Proficient.in.English)

mean(CHR2022PA$X..female)

sd(CHR2022PA$X..female)

mean(CHR2022PA$X..rural)

sd(CHR2022PA$X..rural)

mean(CHR2022PA$Years.of.Potential.Life.Lost.Rate)

sd(CHR2022PA$Years.of.Potential.Life.Lost.Rate)

summary(X..rural)

hist(CHR2022PA$X..Drive.Alone.to.Work)

describe(CHR2022PA$ Presence.of.Water.Violation)

#######################################################################

#Sub-Section 1.2: Simple Linear Regression

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ Average.Daily.PM2.5, data = CHR2022PA)

summary(m2)

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ X..Severe.Housing.Problems, data = CHR2022PA)

summary(m2)

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ X..Drive.Alone.to.Work, data = CHR2022PA)

summary(m2)

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ X..Long.Commute...Drives.Alone, data = CHR2022PA)

summary(m2)

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ Presence.of.Water.Violation, data = CHR2022PA)

summary(m2)

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ Population, data = CHR2022PA)

summary(m2)

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ X..Less.Than.18.Years.of.Age, data = CHR2022PA)

summary(m2)

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ X..65.and.Over, data = CHR2022PA)

summary(m2)

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ X..Black, data = CHR2022PA)

summary(m2)

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ X..American.Indian...Alaska.Native, data = CHR2022PA)

summary(m2)

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ X..Asian, data = CHR2022PA)

summary(m2)

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ X..Native.Hawaiian.Other.Pacific.Islander, data = CHR2022PA)

summary(m2)

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ X..Hispanic, data = CHR2022PA)

summary(m2)

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ X..Non.Hispanic.white, data = CHR2022PA)

summary(m2)

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ X..Not.Proficient.in.English, data = CHR2022PA)

summary(m2)

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ X..female, data = CHR2022PA)

summary(m2)

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ X..rural, data = CHR2022PA)

summary(m2)

##Sub-Section 1.3: Multiple linear regression of physical environment

m5 <- lm(Years.of.Potential.Life.Lost.Rate~ Average.Daily.PM2.5 + X..Severe.Housing.Problems + X..Drive.Alone.to.Work + X..Long.Commute...Drives.Alone + Presence.of.Water.Violation, data = CHR2022PA)

summary(m5)

##Sub-Section 1.3: Multiple linear regression of Demographic measures

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ Population + X..Less.Than.18.Years.of.Age + X..65.and.Over + X..Black + X..American.Indian...Alaska.Native + X..Asian + X..Native.Hawaiian.Other.Pacific.Islander + X..Hispanic + X..Non.Hispanic.white + X..Not.Proficient.in.English + X..female + X..rural, data = CHR2022PA)

summary(m2)

#correlation matrix

cor(CHR2022PA[, c("Years.of.Potential.Life.Lost.Rate", "Population", "X..Less.Than.18.Years.of.Age", "X..65.and.Over", "X..Black", "X..American.Indian...Alaska.Native", "X..Asian", "X..Native.Hawaiian.Other.Pacific.Islander", "X..Hispanic", "X..Non.Hispanic.white", "X..Not.Proficient.in.English", "X..female", "X..rural")])

cor(CHR2022PA[, c("Years.of.Potential.Life.Lost.Rate", "Average.Daily.PM2.5", "X..Severe.Housing.Problems", "X..Drive.Alone.to.Work", "X..Long.Commute...Drives.Alone", "Presence.of.Water.Violation")])

#scatterplot to check relation between response variable with final predictors

scatterplotMatrix(~ Years.of.Potential.Life.Lost.Rate + X..Long.Commute...Drives.Alone + X..Black + X..Asian + X..Hispanic + X..Non.Hispanic.white + X..Not.Proficient.in.English, data = CHR2022PA)

smooth=FALSE, diagonal=FALSE)

scatterplotMatrix(~ Years.of.Potential.Life.Lost.Rate + Population + X..Less.Than.18.Years.of.Age + X..65.and.Over + X..Black + X..American.Indian...Alaska.Native + X..Asian + X..Native.Hawaiian.Other.Pacific.Islander + X..Hispanic + X..Non.Hispanic.white + X..Not.Proficient.in.English + X..female + X..rural, data = CHR2022PA)

smooth=FALSE, diagonal=FALSE)

##forward stepwise regression of physical environment was selected for answer.

m5 <- lm(Years.of.Potential.Life.Lost.Rate~ Average.Daily.PM2.5 + X..Severe.Housing.Problems + X..Drive.Alone.to.Work + X..Long.Commute...Drives.Alone + Presence.of.Water.Violation, data = CHR2022PA)

ols\_step\_forward\_aic(m5)

ols\_step\_forward\_p(m5, penter = 0.05)

#forward stepwise regression of demographic measures

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ Population + X..Less.Than.18.Years.of.Age + X..65.and.Over + X..Black + X..American.Indian...Alaska.Native + X..Asian + X..Native.Hawaiian.Other.Pacific.Islander + X..Hispanic + X..Non.Hispanic.white + X..Not.Proficient.in.English + X..female + X..rural, data = CHR2022PA)

summary(m2)

ols\_step\_forward\_aic(m2)

ols\_step\_forward\_p(m2, penter = 0.05)

#TO find interaction between Years.of.Potential.Life.Lost.Rate ~ X..Long.Commute...Drives.Alone + variables X..Black + X..Asian + X..Hispanic + X..Non.Hispanic.white + X..Not.Proficient.in.English

#final model

m0 <- lm(Years.of.Potential.Life.Lost.Rate~ X..Long.Commute...Drives.Alone + X..Less.Than.18.Years.of.Age, data = CHR2022PA)

summary(m0)

plot(m0)

#final interaction model based on forward stepwise

m\_int <- lm(Years.of.Potential.Life.Lost.Rate~ X..Long.Commute...Drives.Alone + X..Less.Than.18.Years.of.Age + X..Long.Commute...Drives.Alone \* X..Less.Than.18.Years.of.Age, data = CHR2022PA)

summary(m\_int)

plot(m\_int)

#Anova test of the models

m0 <- lm(Years.of.Potential.Life.Lost.Rate~ 1, data = CHR2022PA)

m6 <- update(m0, ~ X..Long.Commute...Drives.Alone, data = CHR2022PA)

m7 <- update(m6, ~. + X..Less.Than.18.Years.of.Age,data = CHR2022PA)

m8 <- update(m7, ~. + X..Long.Commute...Drives.Alone + X..Less.Than.18.Years.of.Age, data = CHR2022PA)

m9 <- update(m8, ~. + X..Long.Commute...Drives.Alone : X..Less.Than.18.Years.of.Age, data = CHR2022PA)

anova(m9,m8)

Anova(m9,m8)

Anova(m8,m7)

Anova(m7,m6)

Anova(m6,m0)

#transformation independently with X..Less.Than.18.Years.of.Age

with(CHR2022PA, invTranPlot(X..Less.Than.18.Years.of.Age, Years.of.Potential.Life.Lost.Rate))

#transformation independently with X..Long.Commute...Drives.Alone

with(CHR2022PA, invTranPlot(X..Long.Commute...Drives.Alone, Years.of.Potential.Life.Lost.Rate))

#residual analysis using residual plot of final model

residualPlots(m0)

#nonconstant variance of final model

ncvTest(m0)

#outlier deterction of interaction model

outlierTest(m0)

#normality assumption q-q plot

qqPlot(m0)