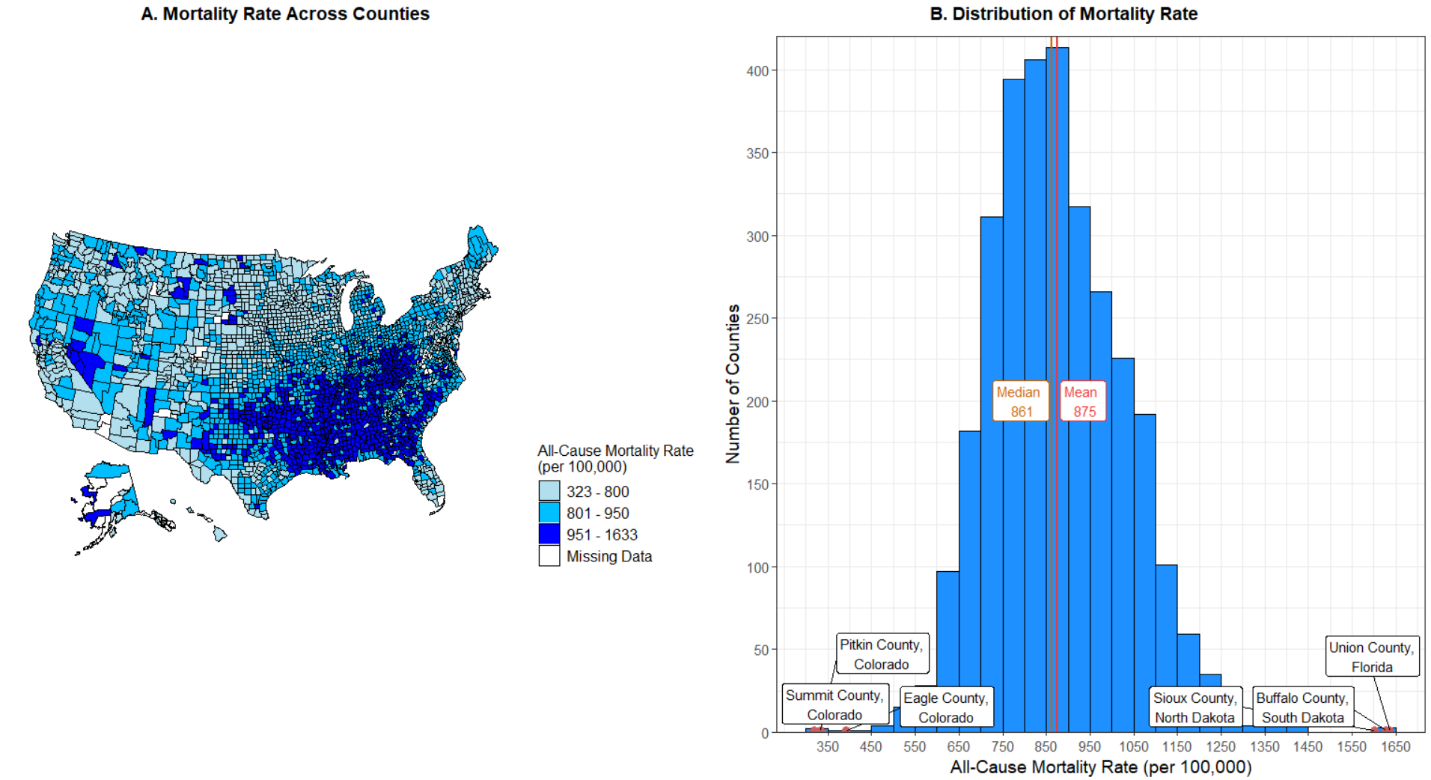
# Quantitative Methods Coursework Assignment

## Question 1

# import tidyverse lib for data manipulation   
library(tidyverse)   
   
# import datasets   
mortality <- read\_csv("county\_mortality.csv") %>%   
 rename(FIPS = fips) # rename from "fips" to "FIPS" for joining   
covariates <- read\_csv("county\_covariates.csv")   
   
# join datasets, removing duplicate columns of statename and countyname   
county\_data <- inner\_join(mortality, covariates, by = "FIPS") %>%   
 select(!c("statename", "countyname"))

## Question 2

# import more libraries for visualisation   
library(ggpubr)   
library(ggrepel)   
library(usmap)  
   
# to find what the ranges of each tertile is   
unique(cut(county\_data$all\_cause,   
 breaks = quantile(county\_data$all\_cause,   
 probs = seq(0,1, by = 0.33)))) # split into tertiles   
   
# create new dataframe to prep data for plotting   
choromap\_data <- county\_data %>%   
 select(c("FIPS", "all\_cause")) %>% # select only the essential columns   
 rename(fips = FIPS, value = all\_cause) %>% # rename for plotting   
 # split the values into tertiles using the ranges found above   
 mutate(tertile = as.factor(   
 ifelse(value <= 800, "1",   
 ifelse(value <= 950, "2", "3"))))   
   
# create first graphic   
g1 <- plot\_usmap(regions = c("county"),   
 data = choromap\_data,   
 values = "tertile") +   
 # set the labels and the colors to make sure it's nice!   
 scale\_fill\_manual(values = c('NA' = "white", "4" = "darkblue", "3" = "blue",   
 "2" = "deepskyblue", "1" = "lightblue2"),   
 labels = c("323 - 800",   
 "801 - 950", "951 - 1633", "Missing Data"),   
 name = "All-Cause Mortality Rate\n(per 100,000)") +   
 # adjust the legend so it doesn't block Hawaii   
 theme(legend.position = "right",   
 legend.text = element\_text(size = 12),   
 legend.title = element\_text(size = 12))   
   
# create simple function to find top 3 mortality rates and bottom 3 mortality rates   
is\_topbottom <- function(x) {   
 return(x < 400 | x > 1500)   
}   
   
# create second graphic   
g2 <- county\_data %>%   
 # create new topbottom value and label columns for plotting   
 mutate(topbot = ifelse(is\_topbottom(all\_cause), paste(county,state,sep=",\n"), as.numeric(NA)),   
 topbot\_values = ifelse(is\_topbottom(all\_cause), all\_cause, as.numeric(NA))) %>%   
 # begin plot   
 ggplot(aes(x=all\_cause)) +   
 # create histogram to show distribution   
 geom\_histogram(fill = "dodgerblue1",   
 color = "black",   
 binwidth = 50,   
 boundary = 0) +   
 # use jitter to plot the top and bottom counties   
 geom\_jitter(aes(x=topbot\_values, y = 1),   
 na.rm = TRUE,   
 color = "indianred",   
 size = 3,   
 alpha = 1.0,   
 position = position\_jitter(width = 5, height = 0, seed = 3)) +   
 # use label\_repel to plot their labels separate from each other   
 geom\_label\_repel(aes(x = topbot\_values, y = 1, label = topbot),   
 box.padding = 0.3,   
 color = "black",   
 position = position\_jitter(width = 5, height = 0, seed = 3)) +   
 # plot mean and median values as lines   
 geom\_segment(aes(y = 0, yend = 420, x = mean(all\_cause), xend = mean(all\_cause)),   
 size = 1, color = "brown2") +   
 geom\_segment(aes(y = 0, yend = 420, x = median(all\_cause), xend = median(all\_cause)),   
 size = 1, color = "darkorange3") +   
 # label the lines, using offsets   
 annotate(geom = "label", x = mean(county\_data$all\_cause)+60, y = 200,   
 label = paste("Mean \n",   
 round(mean(county\_data$all\_cause))),   
 color = "brown2") +   
 annotate(geom = "label", x = median(county\_data$all\_cause)-70, y = 200,   
 label = paste("Median \n",   
 round(median(county\_data$all\_cause))),   
 color = "darkorange3") +   
 scale\_x\_continuous(breaks = seq(350, 1650, 100)) +   
 scale\_y\_continuous(breaks = seq(0, 400, 50), expand = c(0,0)) +   
 labs(title = "",   
 x = "All-Cause Mortality Rate (per 100,000)",   
 y = "Number of Counties") +   
 theme\_bw(base\_size = 14)   
   
  
# use ggarrange to combine both charts into a single visualisation and name them   
g\_combined <- ggpubr::ggarrange(g1, g2, labels = c("A. Mortality Rate Across Counties",   
 "B. Distribution of Mortality Rate"),   
 hjust = c(-0.5, -0.8))   
   
g\_combined

***Figure 1A:*** *Mortality Rate Across Counties. It indicates that there is quite a diversity of mortality rates, from 323 - 1633. The highest mortality rates appear to be clustered in the South and South-East, with a generally lower mortality rate in the Western regions.****Figure 1B:*** *Distribution of Mortality Rate. We see a generally normal distribution, with a mean at 875 and the median just slightly lower at 861. The three counties with the highest and lowest mortality rates have been labelled. It is interesting that the three counties with the lowest mortality rate are all in Colorado, a state in the Mountain West region.*

## Question 3

### **Model Assumptions**

There are numerous assumptions made with linear regression models. I touch on three that I think are most pertinent to this model.

1. The first assumption of the linear regression model is that there is linear relationship between all-cause mortality rates and the independent variables selected.
2. The second assumption is that the independent variables should not be highly correlated (i.e. suffer from multicollinearity). A high correlation between two factors (typically > 0.8) could suggest that both variables represent the same underlying factor and one should be removed [1, p. 101]. In this case, high school education and median income would be the two variables must likely to suffer from this. However, a quick check (cor(county\_data$income\_median, county\_data$edu\_hs)) places correlation at 0.55, which is below the "high correlation" level.
3. A third assumption of the model is that the residuals follow a normal distribution. This is an important assumption because if the residuals are non-normally distributed, it implies that the expected error of the model will vary at different levels of all-cause mortality rates (the dependent variable), making our model less robust.

### **Selection of Variables**

The following variables were selected:

1. Median Household Income: The association between income and life expectancy/mortality is well established across a number of studies [2]. Given its significant correlation with access to medical care and its relation to physical environmental factors, it is likely that it would have an effect on mortality and should hence be retained.
2. High School Education: High School Education or Level of Educational Attainment has also been shown to have strong correlations with mortality, acting as an *upstream* factor that is associated with exposure to other factors (such as social behaviour and income) associated with mortality rates [3].
3. Race (Black) and Race (Others): In the United States, race has been shown to correlate with mortality. In fact, it is not just race but also skin tone that appears to be a significant determinant of mortality [4]. Among blacks for instance, light skinned blacks had the lowest mortality hazards while survey participants with medium to dark brown skin experienced higher mortality hazards [4]. We hence chose to retain this variable.
4. Population Density: Intuitively, population density makes sense to be included as a variable given that it might have effects on intermediate factors (like mental health) that would have an effect on mortality. The relationship between mortality and population density has also been shown in The Netherlands [5] and Japan [6].

The following variables were left out:

1. State and County Name: Geographic location has been shown to play a part in mortality rates [7]-[8]. However, it is not the geographic location per se that affects mortality rates but rather *characteristics* of the location. These characteristics, like income and population density, have been captured in the aforementioned variables and hence geographic locations as a variable have been left out.
2. FIPS: Similar to the State and County Name variable, FIPS is an identifier of geographic location and has similarly been left out.

### **Results**

# create the model   
model <- lm(all\_cause ~ income\_median + edu\_hs + race\_black   
 + race\_other + pop\_density,   
 data = county\_data)   
   
# view output of model   
summary(model)  
confint(model)

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Estimate | 95% CI | p |
| Intercept | 1628 | [1573, 1684] | <0.001 |
| Median Income | -0.005140 | [-0.0005535, 0.0004746] | <0.001 |
| High School Edu | -624.5 | [-695.8, -553.2] | <0.001 |
| Race (Black) | 190.2 | [161.4, 219.0] | <0.001 |
| Race (Others) | 173.1 | [122.0, 224.2] | <0.001 |
| Pop Density | -0.01476 | [-0.02071, -0.008821] | <0.001 |

*Table 1: Summary of Model Intercepts*

We see from the output that all the independent variables are statistically significantly related to all-cause mortality, all having a p-value of <0.001.

#### **Positive Effect**

Median income, high school education, and population density all appear to have a positive effect on all-cause mortality, with a 1 unit increase in the variable leading to a concomitant decrease in mortality. For instance, with every 0.1 (or 10%) increase in proportion of population with high school education, the all-cause mortality rate (per 100,000) is estimated to decrease by ~ 62 deaths (to nearest whole number).

#### **Negative Effect**

Race proportion is shown to have a negative effect on all-cause mortality. Counties that had a higher proportion of non-white races had an increased mortality, with every unit increase in the proportion of non-white races leading to an increase in mortality. For instance, with every 0.1 (or 10%) increase in the proportion of population that is black, the all-cause mortality rate (per 100,000) is estimated to increase by ~19 deaths (to nearest whole number).

#### **Model Uncertainty and Fit**

|  |  |  |  |
| --- | --- | --- | --- |
| RSE | R-Squared | Adj R-Squared | F-Statistic *p-value* |
| 108.5 | 0.4807 | 0.4799 | < 0.001 |

*Table 2: Summary of Model*

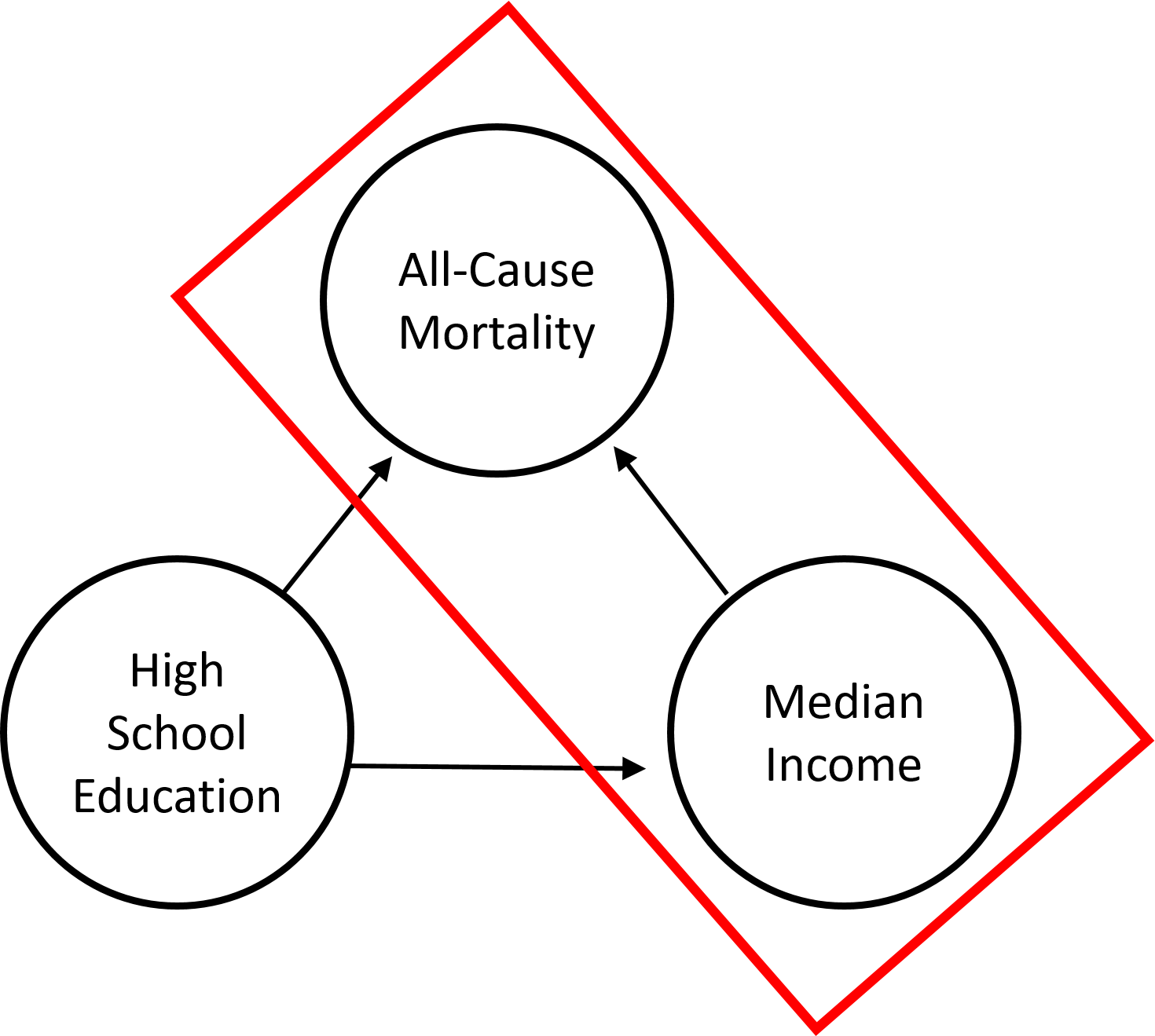
Model uncertainty and fit can be ascertained from the Residual Standard Error (RSE). In our model, for every prediction made about mortality, we can expect, on average, a deviation of 108.5 deaths per 100,000 people. Given the intercept estimate of 1628 and an RSE of 108.5, the percentage error is approximately 6.67%, which in the context of mortality rates appears to be a relatively acceptable level of uncertainty. Model fit can also be assessed by looking at the R-squared value or Adjusted R-squared value (which takes into account the number of parameters to prevent inflating of R-squared if we add more independent variables). In our case, the adjusted R-Squared is 0.4799 which indicates that approximately 48% of the variability in mortality rate can be explained by our independent variables. The evaluation of R-Squared is highly dependent on the context. In this case, given that it is an observational study in the social sciences, we can say that the model is moderately well-fitted to the data with an adjusted R-Squared of 0.4799, following Henseler et al.'s [9] metric of strong (0.75), moderate (0.5), and weak (0.25). Mortality can be affected by such a large variety of factors and is it hence unsurprising that our model, with only 5 parameters, fails to explain a large proportion of the variability in mortality rate between counties. The F-Statistic also gives us an idea of the fit of the model. As it is statistically significant, it implies that the model fits the data better than a model with no independent variables. This means that the inclusion of our independent variables (like income and race) have improved the fit of the model and that together, they are a good fit in explaining the dependent variable.

## Question 4

# create residual model by regressing income\_median against other variables   
income\_model <- lm(income\_median ~ edu\_hs + race\_black + race\_other + pop\_density,   
 data = county\_data)   
   
# add residuals to dataset   
isol\_data <- county\_data %>%   
 mutate(residuals = income\_model$residuals)   
   
# regress residuals against mortality   
residual\_model <- lm(all\_cause ~ residuals, data = isol\_data)   
   
summary(residual\_model)

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Estimate | 95% CI | p |
| Intercept | 874.8 | [869.7, 879.8] | <0.001 |
| Residuals | -0.005140 | [-0.0005535, 0.0004746] | <0.001 |

*Table 3: Summary of Residual Model Intercepts*

The estimates in both models are equal. To understand this more intuitively, we can strip the model to just three variables: all-cause mortality, median income, and high school education. The following figure attempts to depict the relationship between mortality, income, and the covariate (high school education). The red box highlights the effect we are trying to isolate by removing the effect of the covariate.

*Figure 2: A casual graph between all-cause mortality and median income, illustrating the effects of high school education*

In model 1, the estimate for median income shows the relationship between median income and mortality, controlling for high school education:

1. High school education could have an effect on mortality through income - for instance, getting a high school education likely means your median income is higher.
2. We want to get rid of that effect because we want to find the effect of **ONLY income** on mortality. Therefore, we need to remove the effect of high school education on median income and on all-cause mortality.
3. To remove this effect, we need to control for it by holding it constant by setting it to, for instance, its mean. Without high school education changing, we can now see the true effect of income on mortality.
4. in model 1 hence tells us the association between all-cause mortality and median income, after accounting for the variation caused by high school education.

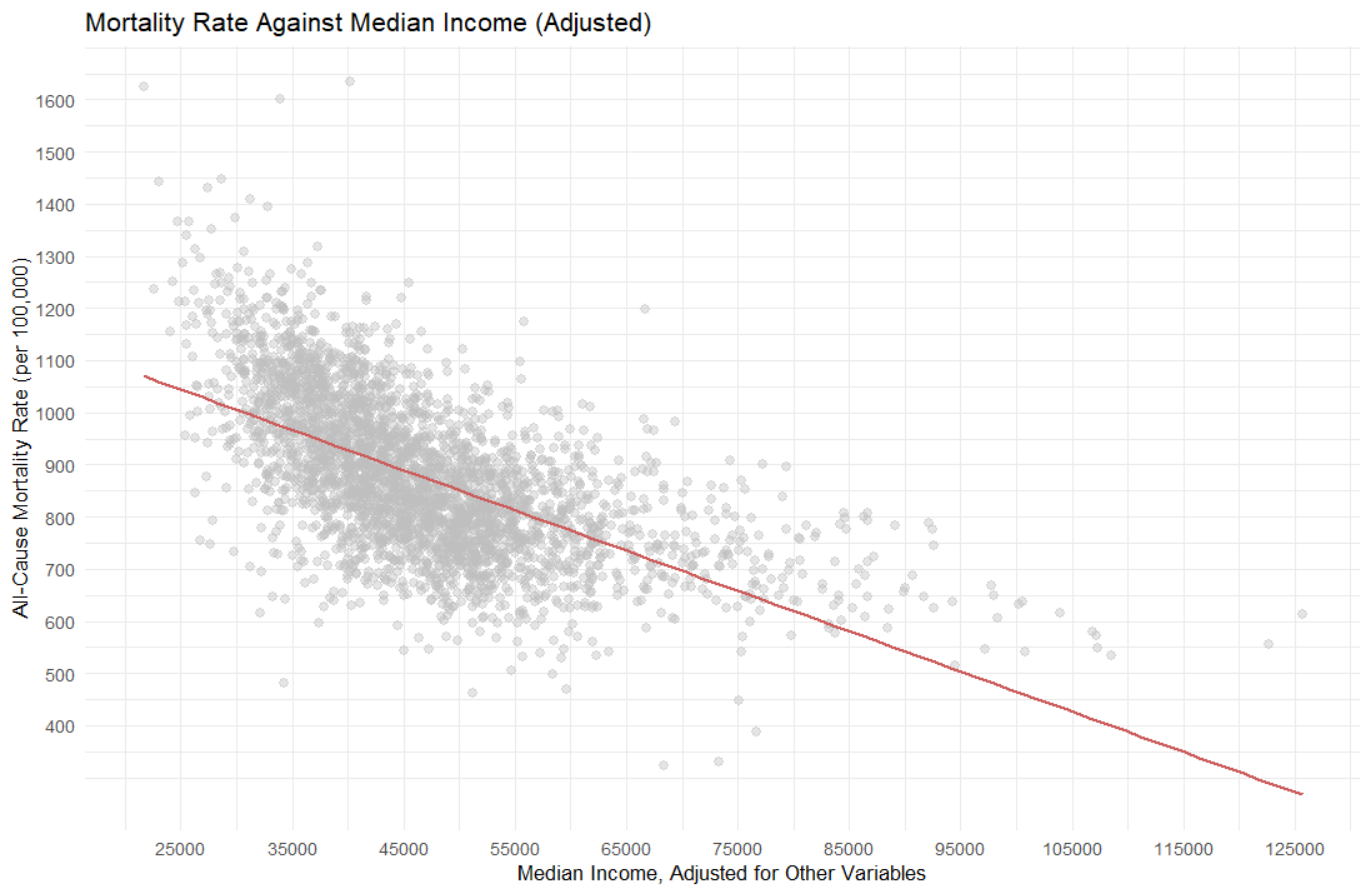
We then regress median income against high school education:

1. The regression model gives us the association between median income and high school education.
2. However, there is some part of median income that cannot be explained by high school education. This is captured in the residuals.
3. Said another way, the residuals captures median income after removing high school education as an explanatory variable.

Therefore, when we regress all-cause mortality against these residuals, we are effectively doing the same thing as we did in model 1. We are finding the effect of **ONLY income** on mortality, after taking into account high school education's effect on income. Expand this to include the rest of the variables (proportion of non-white races and population density) and this explains why in both models is the same.

## Question 5

# import library in order to get data of adjusted income   
library(visreg)

# use visreg to generate data of adjusted income, setting   
# conditions (cond) to use the mean of the other variables   
adjusted <- visreg(model, "income\_median",   
 collapse = FALSE,   
 plot = FALSE,   
 cond = list(edu\_hs = mean(isol\_data$edu\_hs),   
 race\_black = mean(isol\_data$race\_black),   
 race\_other = mean(isol\_data$race\_other),   
 pop\_density = mean(isol\_data$pop\_density)))   
   
# create plot to show variation of mortality with income, after controlling   
# for other covariates   
g3 <- isol\_data %>%   
 ggplot(aes(y = all\_cause, x = adjusted)) +   
 geom\_point(color = "grey", size = 2.5, alpha = 0.4) +   
 geom\_smooth(method = "lm", color = "indianred",   
 se = FALSE) +   
 scale\_y\_continuous(breaks = seq(400, 1600, 100)) +   
 scale\_x\_continuous(breaks = seq(25000, 125000, 10000)) +   
 theme\_minimal() +   
 labs(x = "Median Income, Adjusted for Other Variables",   
 y = "All-Cause Mortality Rate (per 100,000)",   
 title = "Mortality Rate Against Median Income (Adjusted)")   
   
g3

*Figure 3: Mortality Rate Against Median Income (Adjusted). This shows the relationship between median income and all-cause mortality rate, after accounting for the variation caused by other variables (by setting them to their mean). This indicates that increasing median income in a county leads to decreasing all-cause mortality rates*

## Question 6

model\_easy <- lm(all\_cause ~ log(income\_median) + edu\_hs + race\_black   
 + race\_other + pop\_density,   
 data = county\_data)   
   
# get the estimate for log(income\_median)   
coef(model\_easy)["log(income\_median)"]

## log(income\_median)   
## -286.0402

Since we used a log scale, we need to divide the parameter estimate by 100. However, in this case, I chose to divide it by 10 instead to get ~ 29 (to nearest whole number). This implies that for every **10% increase in median income**, the all-cause mortality rate decreases by approximately **29 deaths** per 100,000 people. This is an easier to appreciate metric since (a) a 10% increase in median income is more significant than a $1 increase in income and (b) 29 fewer deaths per 100,000 people is more easily understood than 0.005 deaths.

## Question 7

### **Strengths**

Our first model (model) allows us to clearly appreciate the effects of different variables on mortality rates. Using the estimators, we can easily see how a unit change in the independent variables will affect the mortality rate. The ease of interpretation is a strength of the regression model as this allows it to be easily presented to and understood by a general audience. Another strength of the first model is that it captured the statistically significant association between the explanatory variables and the mortality rate. This would allow us to draw conclusions around how certain socio-economic variables affect the mortality rate within a county. A strength of the second model (model\_easy) is that it condenses the information regarding the other variables, allowing us to focus our attention on the net association between median income and mortality rate. Furthermore, performing a log transformation of median income gave us a better understanding of the relationship between median income in a county and mortality rate, showing us that a 10% increase in median income led to an approximate decrease in the mortality rate of 29 deaths / 100,000 people.

### **Weaknesses**

The first limitation of the models is that is gives us only an idea of the association between the independent variables and mortality rate. However, it does not give us information regarding causality. In order to put into action the findings of the model, we have to better understand the causal relationship between associated variables. For instance, the model shows us that having a higher proportion of non-white residents is associated with a higher mortality rate. More work must be done to elucidate this relationship to ascertain *how* they are related and uncover possible mediating factors. The second limitation of the models is its low R-Squared value. As aforementioned, R-Squared values are highly contextual and hence a low R-Squared value does not always indicate a poor model. However, what it does indicate is that there is much variation in mortality that is unexplained. Other variables such as "health system, social capital and social cohesion variables" [10] could also be examined at the county level. Our models are hence limited by the lack of additional data surrounding mortality rates. A third limitation of the models is that it uses all-cause mortality as the dependent variable. All-cause mortality is useful as it gives a general idea of mortality rates. However, it will not allow us to capture associations between specific causes of death and the independent variables. For instance, a study found an association between specific cancer mortality and population density - lung and kidney cancers were strongly associated with higher population densities while skin cancer was strongly associated with less dense populations [11]. By using all-cause mortality, we are unable to uncover specific trends and create more targeted policies.

### **Conclusions**

All-cause mortality of a county is shown to be strongly associated with median income, proportion of population with high school education, proportion of non-white population, and population density. From this, we can conclude that social and economic factors have a strong relationship with mortality rates. Although we are unable to ascertain direct causal relationships, we can plausibly conclude from past research that these socio-economic factors may be associated with mediating factors. These mediating factors such as access to healthcare [12] and engagement in health-detrimental behaviour (e.g. smoking, drinking) [13] etc. have direct implications on mortality, which may explain the relationship of the model's socio-economic variables with mortality. Another conclusion that we can draw is the specifically strong net association between median income and mortality rate. After controlling for the other variables, we observe the inverse relationship between increasing income and decreasing mortality rate. Again, we can plausibly conclude from past research that this is the case due to the health benefits brought out by having a higher median income. For example, a high income allows access to more nutritious food which has an impact on health and mortality [14]. Lastly, from our initial visualisations, we see (a) large variability in mortality rates across counties in the US and (b) specific geographic patterns with regards to mortality rates. The large variability in mortality rates implies a certain level of socio-economic inequality that exists since, through our models, mortality rates have shown to be associated with median income. This raises a larger question surrounding the issue of inequality and what policies can be put in place to address this. The geographic patterns of generally higher mortality rate in counties in the south and south-west also deserve further investigation as to what characteristics of these regions contribute to this higher mortality rate.

## References

[1] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R: 103*. New York, 2013.

[2] R. Chetty *et al.*, “The Association Between Income and Life Expectancy in the United States, 2001-2014,” *JAMA*, vol. 315, no. 16, p. 1750, Apr. 2016, doi: [10.1001/jama.2016.4226](https://doi.org/10.1001/jama.2016.4226).

[3] R. A. Hummer and E. M. Hernandez, “The Effect of Educational Attainment on Adult Mortality in the United States\*,” *Popul Bull*, vol. 68, no. 1, pp. 1–16, Jun. 2013.

[4] M. R. Benjamins, A. Silva, N. S. Saiyed, and F. G. De Maio, “Comparison of All-Cause Mortality Rates and Inequities Between Black and White Populations Across the 30 Most Populous US Cities,” *JAMA Netw Open*, vol. 4, no. 1, p. e2032086, Jan. 2021, doi: [10.1001/jamanetworkopen.2020.32086](https://doi.org/10.1001/jamanetworkopen.2020.32086).

[5] M. A. Beenackers, J. Oude Groeniger, C. B. M. Kamphuis, and F. J. Van Lenthe, “Urban population density and mortality in a compact Dutch city: 23-year follow-up of the Dutch GLOBE study,” *Health & Place*, vol. 53, pp. 79–85, Sep. 2018, doi: [10.1016/j.healthplace.2018.06.010](https://doi.org/10.1016/j.healthplace.2018.06.010).

[6] T. Nakaya *et al.*, “Associations of All-Cause Mortality with Census-Based Neighbourhood Deprivation and Population Density in Japan: A Multilevel Survival Analysis,” *PLOS ONE*, vol. 9, no. 6, p. e97802, Jun. 2014, doi: [10.1371/journal.pone.0097802](https://doi.org/10.1371/journal.pone.0097802).

[7] R. Subedi, T. L. Greenberg, and S. Roshanafshar, “Does geography matter in mortality? An analysis of potentially avoidable mortality by remoteness index in Canada,” *Health Rep*, vol. 30, no. 5, pp. 3–15, May 2019, doi: [10.25318/82-003-x201900500001-eng](https://doi.org/10.25318/82-003-x201900500001-eng).

[8] Y. C. Vierboom, S. H. Preston, and A. S. Hendi, “Rising geographic inequality in mortality in the United States,” *SSM - Population Health*, vol. 9, p. 100478, Dec. 2019, doi: [10.1016/j.ssmph.2019.100478](https://doi.org/10.1016/j.ssmph.2019.100478).

[9] J. Henseler, C. M. Ringle, and R. R. Sinkovics, “The use of partial least squares path modeling in international marketing,” in *New Challenges to International Marketing*, vol. 20, R. R. Sinkovics and P. N. Ghauri, Eds. Emerald Group Publishing Limited, 2009, pp. 277–319.

[10] P. Baltrus, “Identifying County-Level All-Cause Mortality Rate Trajectories and Their Spatial Distribution Across the United States,” *Prev. Chronic Dis.*, vol. 16, 2019, doi: [10.5888/pcd16.180486](https://doi.org/10.5888/pcd16.180486).

[11] C. Y. Yang and Y. L. Hsieh, “The relationship between population density and cancer mortality in Taiwan,” *Jpn J Cancer Res*, vol. 89, no. 4, pp. 355–360, Apr. 1998, doi: [10.1111/j.1349-7006.1998.tb00571.x](https://doi.org/10.1111/j.1349-7006.1998.tb00571.x).

[12] J. C. Prentice and S. D. Pizer, “Delayed Access to Health Care and Mortality,” *Health Serv Res*, vol. 42, no. 2, pp. 644–662, Apr. 2007, doi: [10.1111/j.1475-6773.2006.00626.x](https://doi.org/10.1111/j.1475-6773.2006.00626.x).

[13] B. K. Payne, J. L. Brown-Iannuzzi, and J. W. Hannay, “Economic inequality increases risk taking,” *Proc Natl Acad Sci USA*, vol. 114, no. 18, pp. 4643–4648, May 2017, doi: [10.1073/pnas.1616453114](https://doi.org/10.1073/pnas.1616453114).

[14] S. A. French, C. C. Tangney, M. M. Crane, Y. Wang, and B. M. Appelhans, “Nutrition quality of food purchases varies by household income: the SHoPPER study,” *BMC Public Health*, vol. 19, no. 1, p. 231, Feb. 2019, doi: [10.1186/s12889-019-6546-2](https://doi.org/10.1186/s12889-019-6546-2).