Final Project

Lorenzo Bartolo

1. Introduction and Statement of Research Question

The relationship between income and health outcomes are well established. However, I wanted to investigate the relationship between an individual's income and their life expectancy. To put simply, does income effect how long you live? I hypothesize that individual income has a significant effect on mortality and life expectancy. The United State's lack of basic social programs, and the privatization of healthcare, simply equates to health outcomes in proportion to an individual's income. Under the classical assumptions of economics, the market for healthcare services would behave identically to any other market, i.e. the market for an automobile; the vehicles with better features, airbags, seatbelts, cupholders, etc., will inherently cost more, and only the individual's who can afford these features will purchase the automobile. This often implies that those with higher incomes can afford vehicles with more advanced safety systems, reducing their chances of being injured or dying in a car crash. I hypothesize that these same market mechanisms apply to healthcare, and are pronounced due to the nature of services being sought. Individual's at the lower end of the income distribution are not able to afford the same healthcare services that the top income earners can afford. It is well established that access to better healthcare improves individual health outcomes and life expectancy. Therefore, it is reasonable to assume that those who do not have access to quality healthcare will have poorer health outcomes and a shorter life expectancy. I test the hypothesis that individual income is a statistically significant indicator on life expectancy and mortality rate. In the model I have included individual income (indv_inc), individual income percentile (indv_pctile), while holding gender (gnd) constant.

Mortality Rate Model Specifications:

$$\widehat{MortalityRate}_{it} = \beta_0 + \beta_1 + u_{it}$$

```
Mortality.model.1 <- lm(data_table_16$mortrate ~ data_table_16$indv_inc)
summary(Mortality.model.1)
coeftest(Mortality.model.1, vcov. = vcovHC, type = "HC1")


Mortality.model.2 <- lm(data_table_16$mortrate ~ data_table_16$indv_inc +
    data_table_16$indv_pctile)
summary(Mortality.model.2)
coeftest(Mortality.model.2, vcov. = vcovHC, type = "HC1")


Mortality.model.3 <- lm(data_table_16$mortrate ~ data_table_16$indv_inc +
    data_table_16$indv_pctile + data_table_16$age_at_d)
summary(Mortality.model.3)
coeftest(Mortality.model.3, vcov. = vcovHC, type = "HC1")
```

Life Expectancy Model Specifications:

$$\widehat{LifeExpectancy}_{it} = \beta_0 + \beta_1 + u_{it}$$

```
Life.Expect.model.1 <- lm(data_table_16$age_at_d ~ data_table_16$indv_inc)
summary(Life.Expect.model.1)
coeftest(Mortality.model.1, vcov. = vcovHC, type = "HC1")

Life.Expect.model.2 <- lm(data_table_16$age_at_d ~ data_table_16$indv_inc +
    data_table_16$indv_pctile)
summary(Life.Expect.model.2)
coeftest(Mortality.model.2, vcov. = vcovHC, type = "HC1")

Life.Expect.model.3 <- lm(data_table_16$age_at_d ~ data_table_16$indv_inc +
    data_table_16$indv_pctile + data_table_16$mortrate)
summary(Life.Expect.model.3)
coeftest(Mortality.model.3, vcov. = vcovHC, type = "HC1")
```

2. Data Description

The data used was deidentified data from federal income tax records from 1999 to 2014. The measure of income was pretax earnings. Mortality and death rates were attained from the Social Security Administration (SSA). There are several limitations to the data: while I would have liked to include more variables in the regression models, such as race or ethnicity, but race and ethnicity are not recorded in tax records. The other limitation within the data is that the data is not made available to the public, so I had to derive the data from a 2016 JAMA study: "The Association Between Income and Life Expectancy in the United States, 2001-2014."

This table reports US mortality rates by gender, age, year and individual income percentile. Individual income percentiles are calculated separately for each gender, age, and year. Incomes are measured two years prior to the mortality rate for mortality rates at ages 40-63, and at age 61 for mortality rates at ages 64-76. The "lag" variable indicates the number of years between measurement of income and mortality. Observations with 1 or 2 deaths have been masked: all mortality rates that reflect only 1 or 2 deaths have been recoded to reflect 3 deaths. Additionally, I added two columns to the dataset with binary "dummy" variables (0 if Male, and 1 if Female) to perform regression on gender.

Table 1: National mortality rates by gender, age, year, and individual income percentile

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
indv_pctile	85,400	50.500	28.866	1	25.8	75.2	100
age_at_d	85,400	55.016	9.246	40	47	62	76
yod	85,400	2,008.033	3.996	2,001	2,005	2,012	2,014
lag	85,400	3.066	2.515	2	2	2	15
mortrate	85,400	0.006	0.006	0.000	0.002	0.008	0.068
indv_inc	85,400	59,746.990	130,190.900	0.486	17,391.660	63,741.860	2,447,709.000
deaths	85,400	77.132	59.199	0	33	106	558
count	85,400	16,145.100	3,992.271	4,730	12,361.8	19,452	23,229
gnd_F	85,400	0.500	0.500	0	0	1	1
gnd_M	85,400	0.500	0.500	0	0	1	1

Table 2:

gnd	indv_pctile	age_at_d	yod	lag	mortrate	indv_inc	deaths	count	gnd_F
F	1	40	2001	2	0.0019118631106013	1.6370008852	40	20922	1
F	1	40	2002	2	0.0026265520534862	1.4019951792	55	20940	1
F	1	40	2003	2	0.0018471709119191	0.9445580756	38	20572	1
F	1	40	2004	2	0.0025641025641026	0.7256385953	51	19890	1
F	1	40	2005	2	0.0025740761926553	0.6136361351	45	17482	1
F	1	40	2006	2	0.0019067915428187	0.5992825352	34	17831	1
F	1	40	2007	2	0.0025565045610365	0.580803894	44	17211	1
F	1	40	2008	2	0.0023925933631541	0.6439602256	46	19226	1
F	1	40	2009	2	0.0021555490824352	0.5465581434	37	17165	1

3. Formulation of the Model

I hypothesize that individual income has a significant effect on mortality and life expectancy. The United State's lack of basic social programs, and the privatization of healthcare, simply equates to health outcomes in proportion to an individual's income. Under the classical assumptions of economics, the market for healthcare services would behave identically to any other market, i.e. the market for an automobile; the vehicles with better features, airbags, seatbelts, cupholders, etc., will inherently cost more, and only the individual's who can afford these features will purchase the automobile. This often implies that those with higher incomes can afford vehicles with more advanced safety systems, reducing their chances of being injured or dying in a car crash. I hypothesize that these same market mechanisms apply to healthcare, and are pronounced due to the nature of services being sought. Individual's at the lower end of the income distribution are not able to afford the same healthcare services that the top income earners can afford. It is well established that access to better healthcare improves individual health outcomes and life expectancy. Therefore, it is reasonable to assume that those who do not have access to quality healthcare will have poorer health outcomes and a shorter life expectancy.

I test the hypothesis that individual income is a statistically significant indicator on life expectancy and mortality rate. In the model I have included individual income (indv_inc), individual income percentile (indv_pctile), while holding gender (gnd) constant.

Mortality Rate Model Specifications:

$$\widehat{MortalityRate}_{it} = \beta_0 + \beta_1 + u_{it}$$

```
Mortality.model.1 <- lm(data_table_16$mortrate ~ data_table_16$indv_inc)
summary(Mortality.model.1)
coeftest(Mortality.model.1, vcov. = vcovHC, type = "HC1")


Mortality.model.2 <- lm(data_table_16$mortrate ~ data_table_16$indv_inc +
    data_table_16$indv_pctile)
summary(Mortality.model.2)
coeftest(Mortality.model.2, vcov. = vcovHC, type = "HC1")


Mortality.model.3 <- lm(data_table_16$mortrate ~ data_table_16$indv_inc +
    data_table_16$indv_pctile + data_table_16$age_at_d + data_table_16$gnd_M)
```

```
summary(Mortality.model.3)
coeftest(Mortality.model.3, vcov. = vcovHC, type = "HC1")
```

Life Expectancy Model Specifications:

$$\widehat{LifeExpectancy}_{it} = \beta_0 + \beta_1 + u_{it}$$

```
Life.Expect.model.1 <- lm(data_table_16$age_at_d ~ data_table_16$indv_inc)
summary(Life.Expect.model.1)
coeftest(Mortality.model.1, vcov. = vcovHC, type = "HC1")

Life.Expect.model.2 <- lm(data_table_16$age_at_d ~ data_table_16$indv_inc +
    data_table_16$indv_pctile)
summary(Life.Expect.model.2)
coeftest(Mortality.model.2, vcov. = vcovHC, type = "HC1")

Life.Expect.model.3 <- lm(data_table_16$age_at_d ~ data_table_16$indv_inc +
    data_table_16$indv_pctile + data_table_16$gnd_F + data_table_16$mortrate)
summary(Life.Expect.model.3)
coeftest(Mortality.model.3, vcov. = vcovHC, type = "HC1")
```

4. Empirical Results

Empirical Results of Mortality Rate Models

The first mortality model (Mortality.model.1), the coefficient representing individual income (indv_inc) was regressed on mortality rates. The average mortality rate for this model was: 6.362e-03. According to the model summary, there is an inverse relationship between income and the mortality rate: on average, as an individual's income rises by 1.00 USD, the individual's mortality rate decreases by -6.7516e-09. The t-value for the individual income coefficient was -32.35. If we used an alpha level of $\alpha = .05$ to determine which predictors were significant in this regression model, we would say that indv_inc is a statistically significant predictor of the mortality rate in this sample. The R-squared for this model was 0.01934, which represents the proportion of the variance for the mortality rate that can be explained by the coefficients. The standard error of the regression line (SER) was 0.006259, which represents the average distance that the observed values of mortality rate fall from the actual values on the regression line. All the coefficients in the models I tested were statistically significant at predicting mortality rates at the level of $\alpha = .05$; Most notably, individual income percentile (indv_pctile) was statistically the most significant indicator of mortality rate (besides age_at_d, which is to be expected) out of all the coefficients, with a t-value of -182.718, which was significant at the level of $\alpha = .001$. These results are summarized in Table 3.

Empirical Results of Life Expectancy Models:

The second model specification was regressing different coefficients on life expectancy. The average age at death (age_at_d) for this model was: 39.97. If we used an alpha level of $\alpha = .05$ to determine which coefficients were significant at predicting life expectancy or age at death, this regression model, we would say that all of the coefficients were statistically significant predictors of life expectancy in this sample. Behind mortality rate, which is to be expected, the most statistically significant coefficient for this regression model was also individual income percentile (indv_pctile). With a t-value of 140.24, it was significant at predicting the life expectancy of individuals at levels down to $\alpha = .001$. The R-squared for this life expectancy model

was 0.6748, which represents the proportion of the variance for life expectancy that can be explained by the coefficients in the model. The standard error of the regression line (SER) was 5.273, which represents the average distance that the observed values of life expectancy fell from the actual values on the regression line. These results are summarized in Table 4.

Table 3: Mortality Models

	Dependent variable:		
	mortrate		
	(1)	(2)	(3)
indv_inc	−0.00000***	0.00000***	−0.00000***
	(0.00000)	(0.00000)	(0.00000)
indv_pctile		−0.00008***	−0.00008***
		(0.000001)	(0.0000004)
age_at_d			0.00051***
			(0.000001)
gnd_M			0.00275***
			(0.00002)
Constant	0.00636***	0.01005***	−0.01938***
	(0.00002)	(0.00004)	(0.00007)
Observations	85,400	85,400	85,400
R^2	0.01934	0.13378	0.73328
Adjusted R^2	0.01933	0.13376	0.73327
Residual Std. Error	0.00626 (df = 85398)	0.00588 (df = 85397)	0.00326 (df = 85395)
F Statistic	1,684.14900*** (df = 1; 85398)	6,594.45300*** (df = 2; 85397)	58,693.17000*** (df = 4; 85395)
Note:			*p<0.1; **p<0.05; ***p<0.01

5. Summary and Discussion

While the results of the regression models were significant, they are not completely surprising. I previously hypothesized that income would be a strong determining factor on health outcomes in the United States, for the reason that healthcare services are bought and sold on a marketplace, and the healthcare services of higher quality will often cost more than those of lesser quality. Consequently, those who have the ability to purchase higher quality healthcare services would realize the benefits as longer life expectancies and lower mortality rates. It is also important to realize that at some income levels, certain healthcare procedures that are not covered by individual's health insurance, would simply be unattainable. To use an exaggerated example, a patient seeking an additional round of chemotherapy treatment for cancer, might not be able to cover the cost of the procedure if it was not covered under their health insurance. Thus, forgoing the additional round of chemotherapy that could have hypothetically extended their life.

One unfortunate limitation of this study, was not being able to accurately control for race or ethnicity. As discussed previously, this information is not recorded on collected tax data, and I would hypothesize that it would also be a significant indicator for life expectancy and mortality rates. Given more time, I would have also liked to compare the results of these models to similar studies done internationally, to contrast the

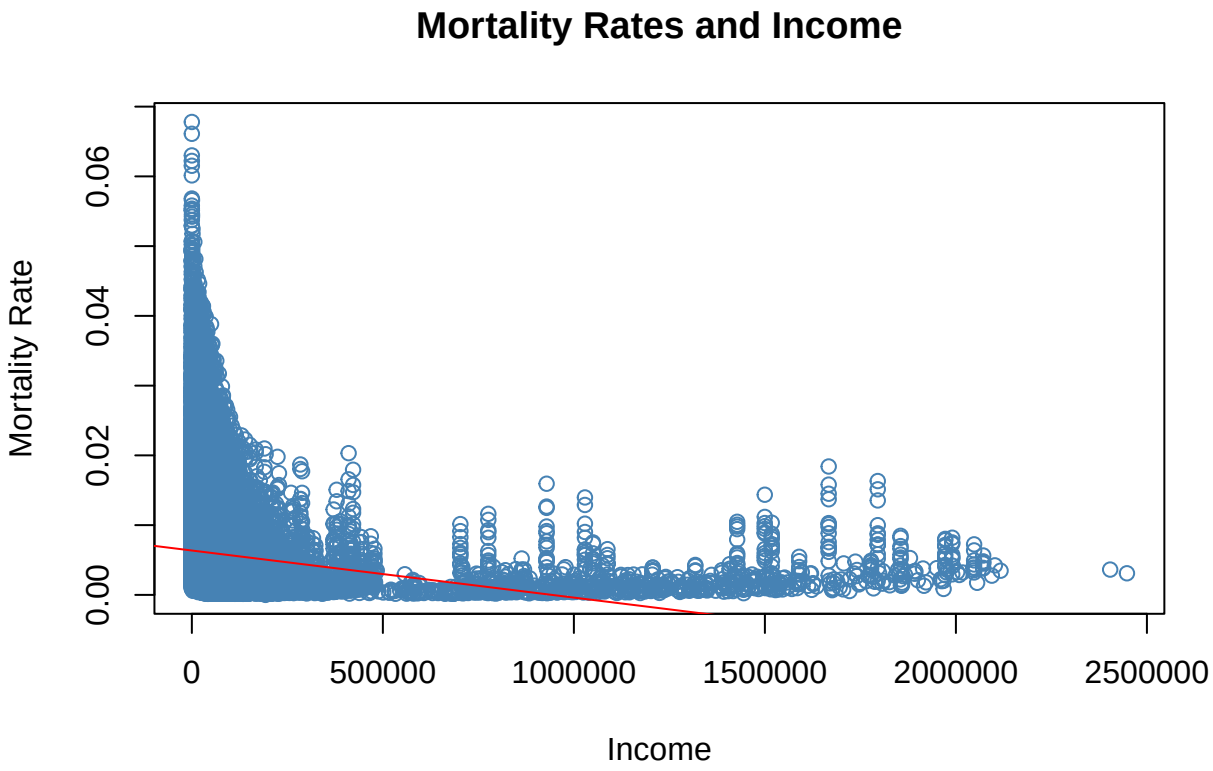
Table 4: Life Expectancy Models

	Dependent variable:		
	age_at_d		
	(1)	(2)	(3)
indv_inc	0.000001***	0.000001***	0.000001***
	(0.0000002)	(0.0000003)	(0.0000002)
indv_pctile		−0.00208*	0.10411***
		(0.00122)	(0.00074)
gnd_F			3.65153***
			(0.03732)
mortrate			1,326.82400***
			(3.15221)
Constant	54.96650***	55.05926***	39.97033***
	(0.03481)	(0.06473)	(0.05431)
Observations	85,400	85,400	85,400
R^2	0.00014	0.00017	0.67483
Adjusted R^2	0.00013	0.00015	0.67481
Residual Std. Error	9.24563 (df = 85398)	9.24553 (df = 85397)	5.27268 (df = 85395)
F Statistic	11.80997*** (df = 1; 85398)	7.34998*** (df = 2; 85397)	44,304.70000*** (df = 4; 85395)
Note:			*p<0.1; **p<0.05; ***p<0.01

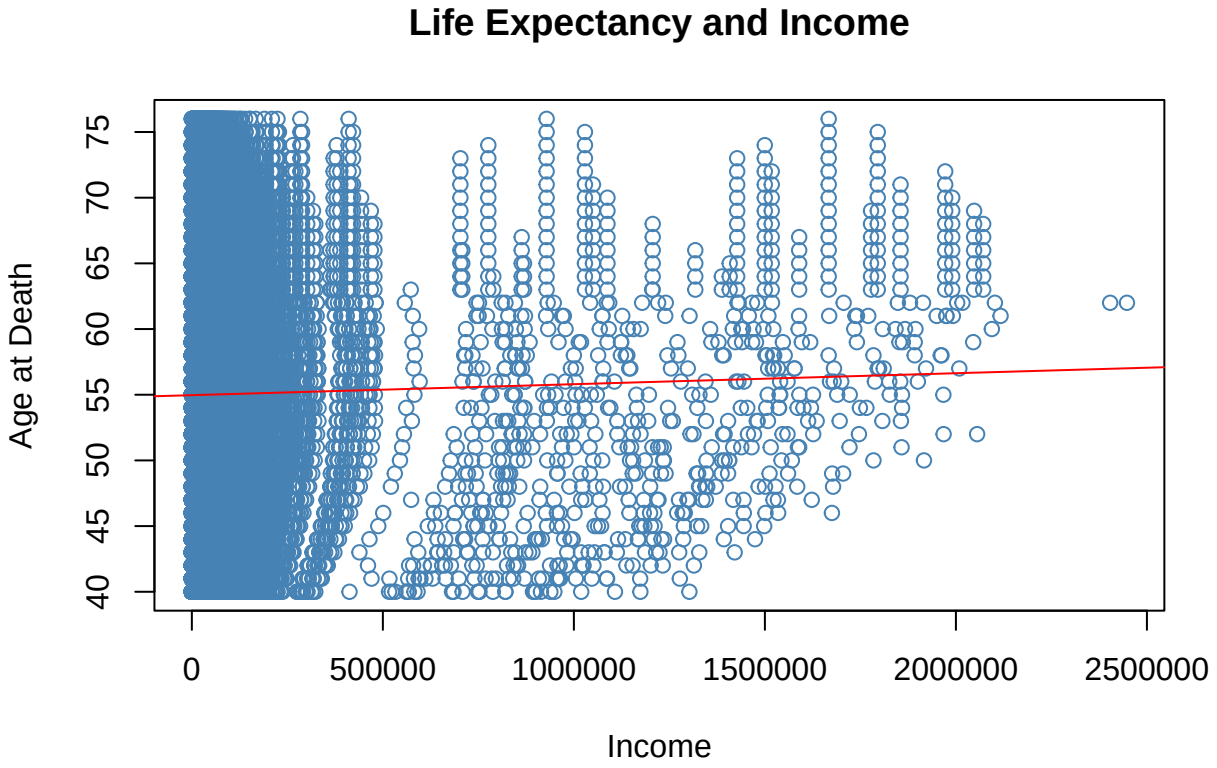
differing results. Additionally, I would like to see how recent COVID-19 data would effect the results of the models.

6. Graphs and Figures

Income and Mortality Rate



Income and Life Expectancy



REFERENCES

- Econometrics with R: https://www.econometrics-with-r.org/10-rwpd.html
- Investopedia: https://www.investopedia.com/terms/r/residual-standard-deviation.asp
- Website: https://www.statology.org/significance-codes-in-r/
- Website: https://www.marsja.se/create-dummy-variables-in-r/
- 2016 JAMA Study: https://jamanetwork.com/journals/jama/fullarticle/2513561?guestAccessKey=4023ce75-d0fb-44de-bb6c-8a10a30a6173
- Introduction to Econometrics Book

Chang, W., J. Cheng, JJ. Allaire, Y. Xie, and J. McPherson. 2015. "Shiny: Web Application Framework for r. R Package Version 0.12.1." Computer Program. http://CRAN.R-project.org/package=shiny.

Stock and Watson 2015. 2019. "Introduction to Econometrics." Journal Article. http://www.R-project.org.