Causal Relationship between Death Rate due to Cancer and Socio-economic Factors

Name: Ritwika Acharjee

Roll No.: C93_ECO_211152

Paper: Econometrics

2nd Semester Internal Examination 2022

Socioeconomic patterns in all-cancer, lung, and colorectal cancer mortality changed dramatically over time. Individuals in more deprived areas or lower education and income groups had higher mortality and incidence rates than their more affluent counterparts, with excess risk being particularly marked for lung, colorectal, cervical, stomach, and liver cancer. Education and income inequalities in mortality from all-cancers, lung, prostate, and cervical cancer increased during 1979–2011. Socioeconomic inequalities in cancer mortality widened as mortality in lower socioeconomic groups/areas declined more slowly.

The Model:

$$target\_deathrate_i = \alpha + \beta_1 incidencerate_i + \beta_2 medincome_i + \beta_3 popest2015_i + \beta_4 povertypercent_i + \beta_5 medianagefemale_i + \beta_6 avghouseholdsize_i + \beta_7 pctemployed16\_over_i + u_i$$

We will estimate this linear regression model. Here we have considered target_deathrate as the dependent variable. The independent variables are avghouseholdsize, incidencerate, medianincome, popest2015, povertypercent, medianagefemale, and pctemployed16_over. Now, our main objective is to explain the variability of target_deathrate for the variability of the independent variables. Basically how the variables can explain the variability of the dependent variable.

For that, we are estimating a Multiple Linear Regression Model. As, the parameters are in linear order so, this model is called a linear model. We will use OLS method to estimate the model.

Data Dictionary-

target_deathRate: Dependent variable. Mean *per capita* (100,000) cancer mortalities

incidenceRate: Mean *per capita* (100,000) cancer diagnoses

medianIncome: Median income per county

popEst2015: Population of county

povertyPercent: Percent of populace in poverty

MedianAgeFemale: Median age of female county residents

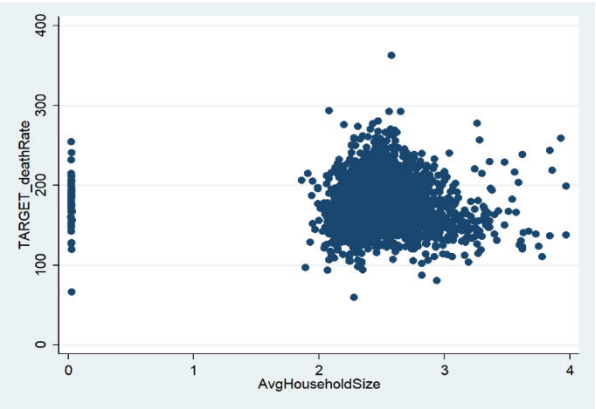
AvgHouseholdSize: Mean household size of county

[*Source:Data.World*]

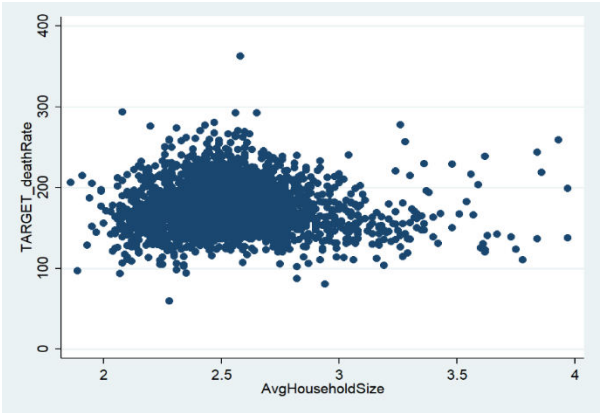
First, we will see the overview of the variables like whether it has any outliers, missing values etc.

Visualization of Data:

Average Household Size-

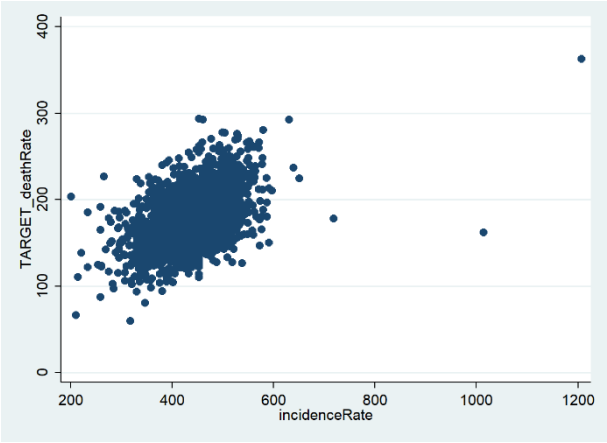


Clearly from the scatter plot, we can tell that average household size can't be less than 1. So those are outliers.

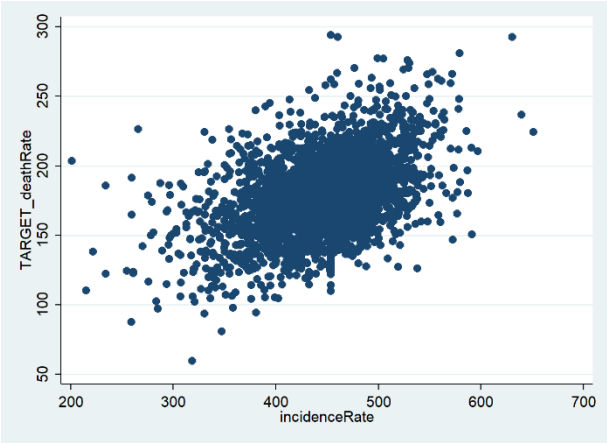


After removing the outliers, this is the plot. And here can we see a weak positive relationship.

Incidence Rate-

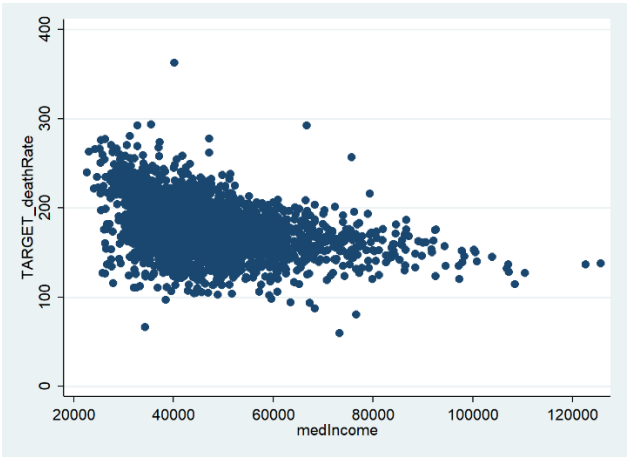


Clearly from the scatter plot, we can tell that there are some incidence rate which take extreme values. So those are outliers.



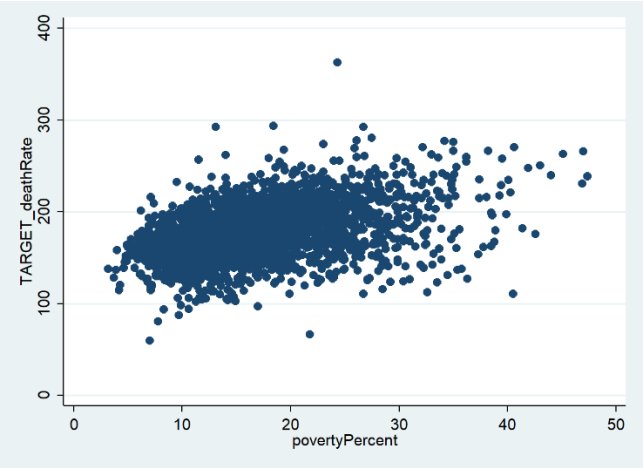
After removing the outliers, this is the plot. And here can we see a positive relationship.

Median Income-



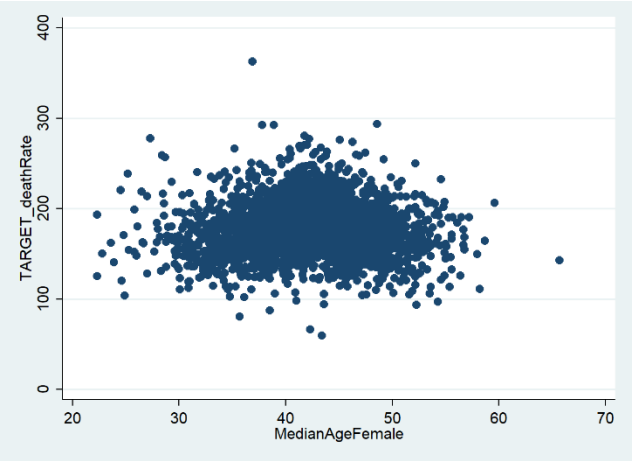
Here can we see a negative relationship. This implies that as the median income rises, death rate falls as people can invest more on their health.

Poverty Percent-



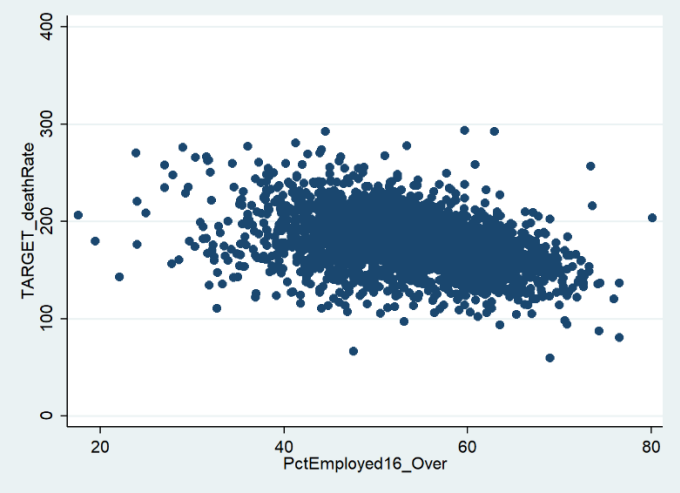
Here can we see a positive relationship which means that, with the increase in poverty per cent among populace, people have less money in their hand to treat diseases like cancer.

Median Female Age-



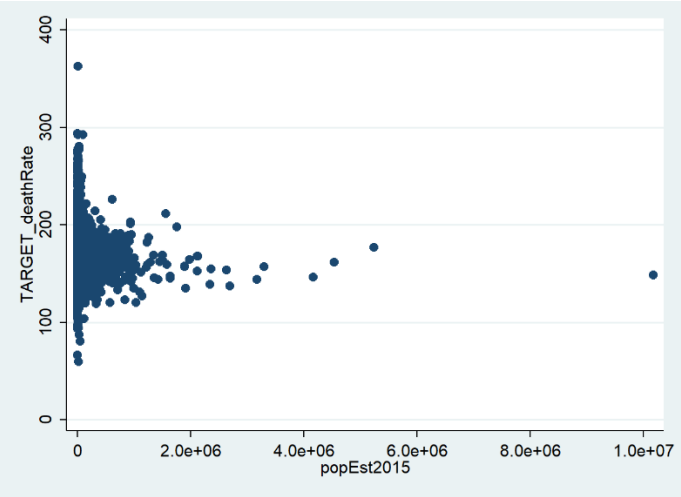
Here can we see no such relationship. There is not much effect between median age of female and target death rate.

Employed Percent over age 16-



Here can we see a negative relationship. As people start becoming employed after 16 years of age, they have more money for consumption and expenditure of heath, and this decreases the death rate of cancer.

Population-



Here can we see no such relationship.

Model Estimation:

We have estimated the model using OLS method. Here is ANOVA table.

Source	SS	df	MS		
				Number of obs	= 2,837
				F(7, 2829)	= 292.63
Model	894404.284	7	127772.041	Prob > F	= 0.0000
Residual	1235248.16	2,829	436.637738	R-squared	= 0.4200
				Adj R-squared	= 0.4185
Total	2129652.44	2,836	750.935276	Root MSE	= 20.896

target_deathrate	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
incidencerate	.2373966	.0078253	30.34	0.000	.2220527	.2527405
medincome	-.0004959	.0000625	-7.94	0.000	-.0006184	-.0003734
popest2015	-5.18e-06	1.22e-06	-4.24	0.000	-7.57e-06	-2.78e-06
povertypercent	.3595665	.1390197	2.59	0.010	.0869763	.6321567
medianagefemale	-.3586861	.1198853	-2.99	0.003	-.5937574	-.1236147
avghouseholdsize	1.926979	2.252388	0.86	0.392	-2.489509	6.343467
pctemployed16_over	-.7020953	.0871985	-8.05	0.000	-.8730744	-.5311162
_cons	138.4255	14.89516	9.29	0.000	109.219	167.6319

Here, we can see the estimated result.

Now, the coefficients we have got we have to check for the significance of the result. For that we have to conduct some tests.

t-test:

Testing the significance of the individual variables.

$$H_0: \beta_i = 0$$

$$H_1: \beta_i \neq 0$$

The $t-$ *statistic for this* $t-$ *test is given by*:

$$t = \frac{\hat{\beta}_i - 0}{se(\hat{\beta}_i)}$$

All the variables expect averagehousehold is significant at 5% level as the p values are very low for those variables. Here, we can see the t-statistic for the variables and after conducting the t-test, we can find that average household size is not significant.

F-test:

Testing for significance of the overall fit of the model.

$$H_0: \beta_1 = \beta_2 \dots \dots \dots = \beta_6 = 0$$

$$H_1: At\ least\ on\ of\ the\ \beta_i \neq 0$$

$The\ F-statistic\ for\ this\ F-test\ is\ given\ by$:

$$F(k, n-k-1) = \frac{\frac{R^2}{k}}{\frac{1-R^2}{n-k-1}}$$

Also here the F-value is significantly high. So, we can conclude that here we reject the null hypothesis that each coefficient is zero. That implies at least one of the coefficients is not equal to zero.

Interpretation:

Incidence Rate-

$$\frac{\partial(target\_deathrate)}{\partial(incidencerate)} = \beta_1 = 0.2373966$$

Here, β_1 is positive. So, it implies that there is a positive relationship between incidence rate and death rate. It implies as the Mean *per capita* (100,000) cancer diagnoses increases the target death rate increases and as there are more diagnoses of cancer, there will be more death. That implies for one unit increase in incidence rate, the target death rate increases by 0.2259501 units.

Median Income-

$$\frac{\partial(target\_deathrate)}{\partial(medincome)} = \beta_2 = -0.0004959$$

Here, β_2 is negative. So, it implies that there is a negative relationship between median income and death rate. It implies as the median income of a region increases the target death rate falls and it is quite obvious as more income leads to better access to health and other things. That implies for one unit increase in median income, the target death rate falls by 0.0004942 unit.

Population-

$$\frac{\partial(target\_deathrate)}{\partial(popest2015)} = \beta_3 = -5.04e-06$$

The β_3 value is infinitesimally small; therefore it has no effect on the dependent variable.

Poverty per cent-

$$\frac{\partial(target\_deathrate)}{\partial(povertypercent)} = \beta_4 = .415972$$

The value of β_4 is positive. This implies that one unit increase in poverty per cent, target death rate increases by .415972 units. This socio-economic factor has a great impact in the increase of cancer mortality rate. Poor people don't have enough money to afford treatments for diseases like cancer, as it is very expensive.

Median Female Age-

$$\frac{\partial(target\_deathrate)}{\partial(medianagefemale)} = \beta_5 = -.2635555$$

The negative relationship between median age of female and target death rate shows that for one unit increase in median age of female, death rate decreases by -.2635555 units. As women attain higher age, they take good care of their health and become mature which leads to better introspection of catching any disease.

Average Household Size-

$$\frac{\partial(target\_deathrate)}{\partial(avghouseholdsize)} = \beta_6 = 2.747482$$

The value of β_6 is positive which implies that one unit increase in average household size, target death rate due to cancer increases by 2.747482 units. With larger household size, it becomes unaffordable to make enough consumption expenditure to fulfil the basic needs like healthy food, sanitation, medicine, etc. Such households tend to get more unwell.

Per cent Employed Over 16 years age-

$$\frac{\partial(target\_deathrate)}{\partial(pctemployed16\_over)} = \beta_7 = -.6490928$$

The β_7 value is negative which implies that as there is more employment after the age of 16, people have greater disposable income that helps them to invest in health and make life affordable. For one unit increase in per cent employed over 16 years of age, target death rate decreases by $-.6490928$.

Checking for assumptions:

Homoscedasticity-

We have to check whether our model is homoscedastic or not i.e. whether the error variance is constant or not. For that we have to conduct two tests; Breusch-Pagan test and White test.

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of target_deathrate

         chi2(1)      =     0.42
         Prob > chi2  =   0.5166
```

From Breusch-Pagan test we can see that the null hypothesis i.e. homoscedasticity is failed to reject. So, by this test our model is homoscedastic.

```
White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(35)     =    340.66
         Prob > chi2  =    0.0000

Cameron & Trivedi's decomposition of IM-test
```

Source	chi2	df	p
Heteroskedasticity	340.66	35	0.0000
Skewness	23.27	7	0.0015
Kurtosis	9.05	1	0.0026
Total	372.98	43	0.0000

As BP test is good for linear relationship, it sometimes doesn't give correct results due to its assumption. We can see that by White's test, we are rejecting the null hypothesis. So, our model violates the assumption of homoscedasticity. So, we will use robust standard error for the estimates.

Linear regression

Number of obs	= 2,837
F(7, 2829)	= 209.16
Prob > F	= 0.0000
R-squared	= 0.4200
Root MSE	= 20.896

target_deathrate	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
incidencerate	.2373966	.0098456	24.11	0.000	.2180913	.2567019
medincome	-.0004959	.0000651	-7.62	0.000	-.0006235	-.0003683
popest2015	-5.18e-06	1.44e-06	-3.59	0.000	-8.00e-06	-2.35e-06
povertypercent	.3595665	.1666987	2.16	0.031	.0327032	.6864298
medianagefemale	-.3586861	.1359741	-2.64	0.008	-.6253044	-.0920677
avghouseholdsize	1.926979	3.011461	0.64	0.522	-3.977903	7.83186
pctemployed16_over	-.7020953	.1064347	-6.60	0.000	-.9107928	-.4933978
_cons	138.4255	18.0828	7.66	0.000	102.9687	173.8823

Multicollinearity-

To check for multicollinearity, we have to check the VIF.

Variable	VIF	1/VIF
povertyper~t	5.21	0.191856
medincome	3.71	0.269487
pctemploye~r	3.43	0.291559
medianagef~e	2.60	0.384769
avghouseho~e	2.01	0.497727
popest2015	1.11	0.897087
incidencer~e	1.05	0.956420
Mean VIF	2.73	

From the table we can see that expect povertypercent, all the variables have very low VIF. VIF>10 is indication of serious multicollinearity problem. But, in our model we can see that multicollinearity is not a serious issue here.

Conclusion-

From this result derived from the regression model, we can conclude that median income per county, and per cent employed over 16 years of age have a positive impact on decreasing the death rate due to cancer. If the household have better income and employment opportunities, then there are less chances of cancer mortality in the economy. When people are poor and the average household size is large, it is not possible to cater the basic amenities and needs of all the members of the household, like health and education, therefore, fewer diseases are diagnosed and less investment is made to improve health.

Reference-

1. Das, P. (2019), Econometrics in Theory and Practice - Analysis of Cross Section, Time Series and Panel Data with Stata 15.1, Singapore, Springer