BIOL 4030 RESEARCH DESIGN and DATA ANALYSIS

THE EFFECT OF SOCIOECONOMIC FACTORS ON CANCER MORTALITY RATES ACROSS COUNTIES IN THE UNITED STATES

A Statistical Analysis

By Pearline N

INTRODUCTION

CANCER: common, prevalent, unpreventable disease

- Treatments available, but **high cost, side effects, mental stress.**
- **Access to treatments -** depends on many socioeconomic factors.
- Factors that may affect the incidence of cancer and the subsequent mortality rate:
  geography + economic status + exposure to carcinogens + poverty levels + access to healthcare etc...
- **To what extent?**

DATA source: "cancer.csv" obtained from previous course.

- 3047 rows represent **different counties across continental states in the United States.**
- 34 columns represent different **socio-economic and health factors** that affect the economic conditions of each county.
- [I subsetted 5 variables from this dataset]

RESEARCH QUESTION

Do socioeconomic factors affect the rate of death due to cancer?
And if they do, is it a positive effect or a negative effect?

Cancer Death Rate = Poverty Percentage, Cancer Incidence Rate, Median Income, Private health insurance

(dependent variable / y)

MULTIPLE LINEAR REGRESSION MODEL

(independent variables / x)

DESCRIPTIVE STATISTICS

```{r}
# we use read.csv() to extract our dataset and store it in a vector.
cancer <- read.csv("cancer.csv")

# we subset our variables of interest into vectors using $
death_rate <- cancer$TARGET_deathRate
med_income <- cancer$medIncome
incidence_rate <- cancer$incidenceRate
pov_percent <- cancer$povertyPercent
pvt_coverage <- cancer$PctPrivateCoverage
summary(cancer[, c(3, 4, 5, 7, 27)])
```

```
 TARGET_deathRate incidenceRate     medIncome      povertyPercent  PctPublicCoverage
 Min.   : 59.7    Min.   : 201.3   Min.   : 22640   Min.   : 3.20   Min.   :11.20
 1st Qu.:161.2    1st Qu.: 420.3   1st Qu.: 38882   1st Qu.:12.15   1st Qu.:30.90
 Median :178.1    Median : 453.5   Median : 45207   Median :15.90   Median :36.30
 Mean   :178.7    Mean   : 448.3   Mean   : 47063   Mean   :16.88   Mean   :36.25
 3rd Qu.:195.2    3rd Qu.: 480.9   3rd Qu.: 52492   3rd Qu.:20.40   3rd Qu.:41.55
 Max.   :362.8    Max.   :1206.9   Max.   :125635   Max.   :47.40   Max.   :65.10
```

```{r}
# we load the dplyr package using library()
# we calculate mean, sd, se and IC of the variable: death rate
library(dplyr)
deathrate <- cancer %>%
  group_by() %>%
  summarise(
    n=n(),
    mean=mean(death_rate),
    sd=sd(death_rate)
  ) %>%
  mutate( se=sd/sqrt(n))  %>%
  mutate( ic=se * qt((1-0.05)/2 + .5, n-1))
deathrate
```

n <int>	mean <dbl>	sd <dbl>	se <dbl>	ic <dbl>
3047	178.6641	27.75151	0.5027481	0.9857598

n <int>	mean <dbl>	sd <dbl>	se <dbl>	ic <dbl>
3047	16.87818	6.409087	0.1161074	0.2276568

n <int>	mean <dbl>	sd <dbl>	se <dbl>	ic <dbl>
3047	448.2686	54.56073	0.9884256	1.938049

n <int>	mean <dbl>	sd <dbl>	se <dbl>	ic <dbl>
3047	47063.28	12040.09	218.119	427.6754

n <int>	mean <dbl>	sd <dbl>	se <dbl>	ic <dbl>
3047	64.35494	10.64706	0.1928827	0.3781935

ASSUMPTIONS TESTED

1. **RANDOMNESS: The data was collected randomly from populations in each county.**
2. **INDEPENDENCE: Each variable was tested separately so they are independent of each other.**
3. **MULTICOLLINEARITY: Are the dependant and independant variables linearly related?**

```{r}
# multiple linear regression model:
cancermodel <- lm(death_rate ~ med_income + incidence_rate + pvt_coverage + pov_percent)
# we load the car package from the library
# we use vif() function: variance inflation factors to check for multicollinearity
library(car)
vif(cancermodel)
```

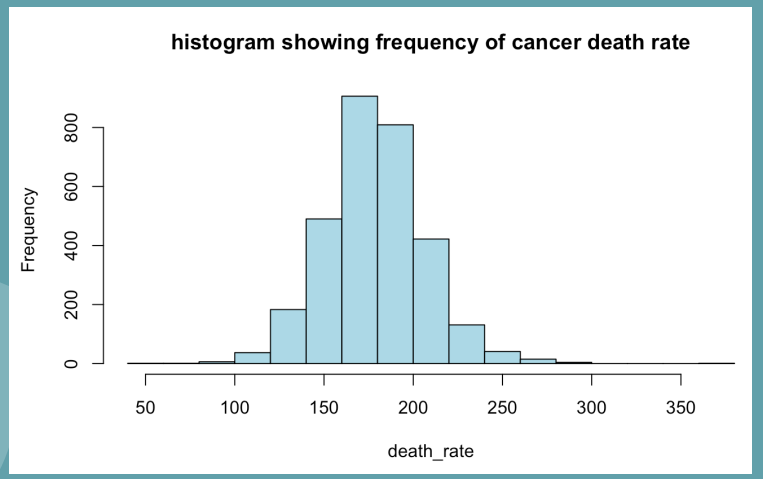
```
med_income incidence_rate   pvt_coverage    pov_percent
  2.780725       1.042116       3.377958       4.155492
```

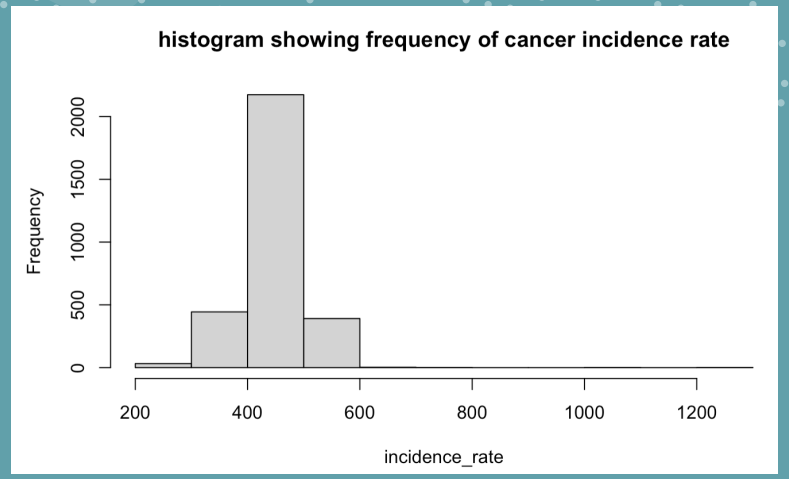
- **Multicollinearity is a condition in which two or more predictor variables in a multiple regression model are highly correlated.**
- **VIF value <= 4 means no multicollinearity (Analytics Vidhya, 2023). The variance of the coefficient is not inflated at all.**
- **From these results, since all the vif values are less than/equal to 4, there is no multicollinearity, thus we can keep the variables.**

https://www.analyticsvidhya.com/blog/2016/07/deeper-regression-analysis-assumptions-plots-solutions/

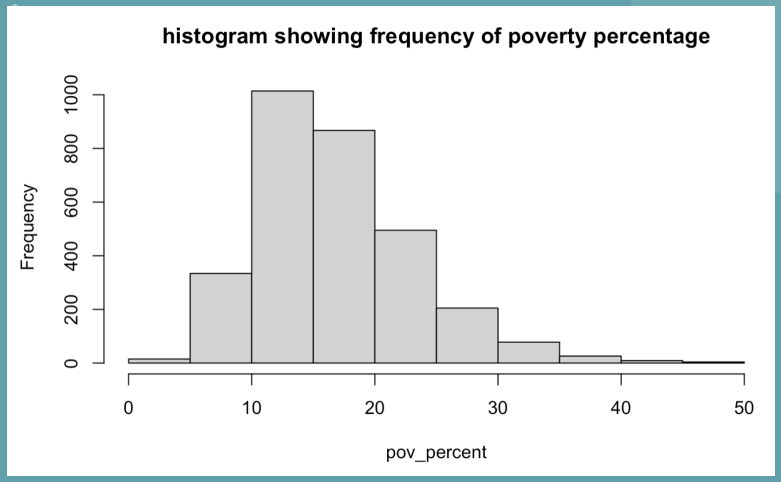
ASSUMPTIONS TESTED

4. NORMALITY: Is the data and residuals normally distributed?

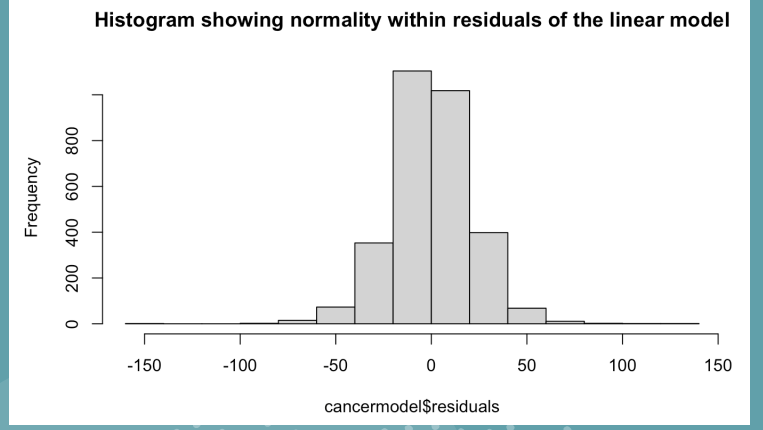


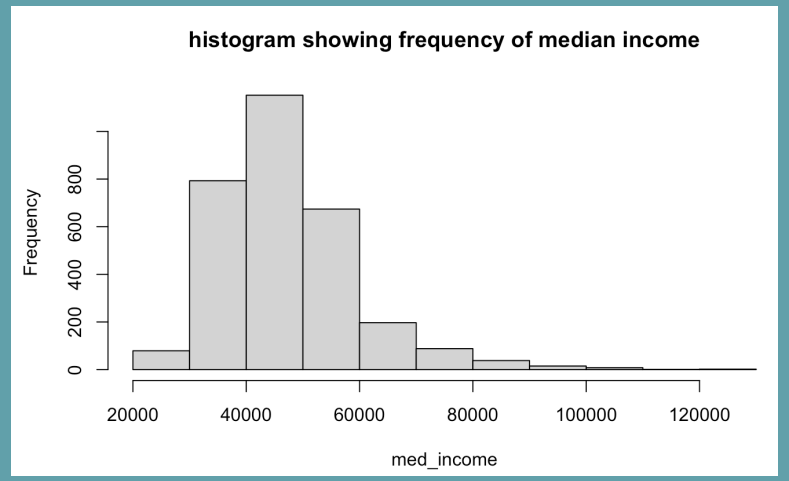


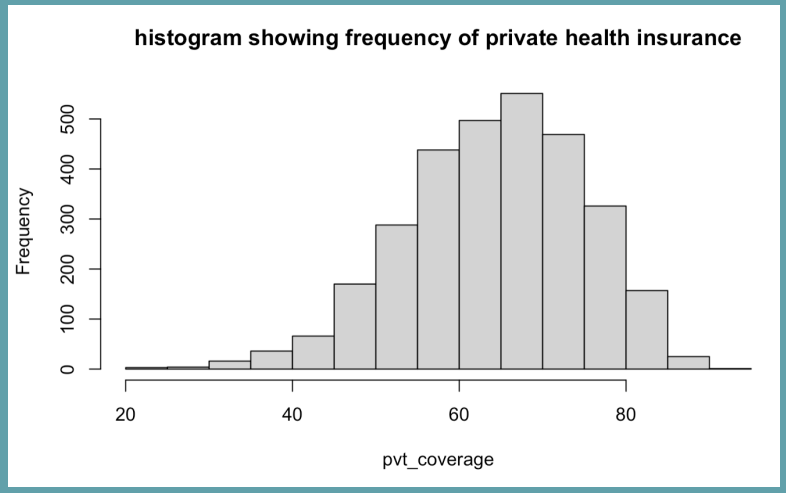

```r
# we use hist() to create histograms of each variable
hist(death_rate, col = "lightblue")
hist(incidence_rate)
hist(med_income)
hist(pov_percent)
hist(pvt_coverage)
```




```r
# we create a histogram of the residuals of our linear model.
hist(cancermodel$residuals, col = "lightgrey",
     main = "Histogram showing normality within residuals of the linear model")
```





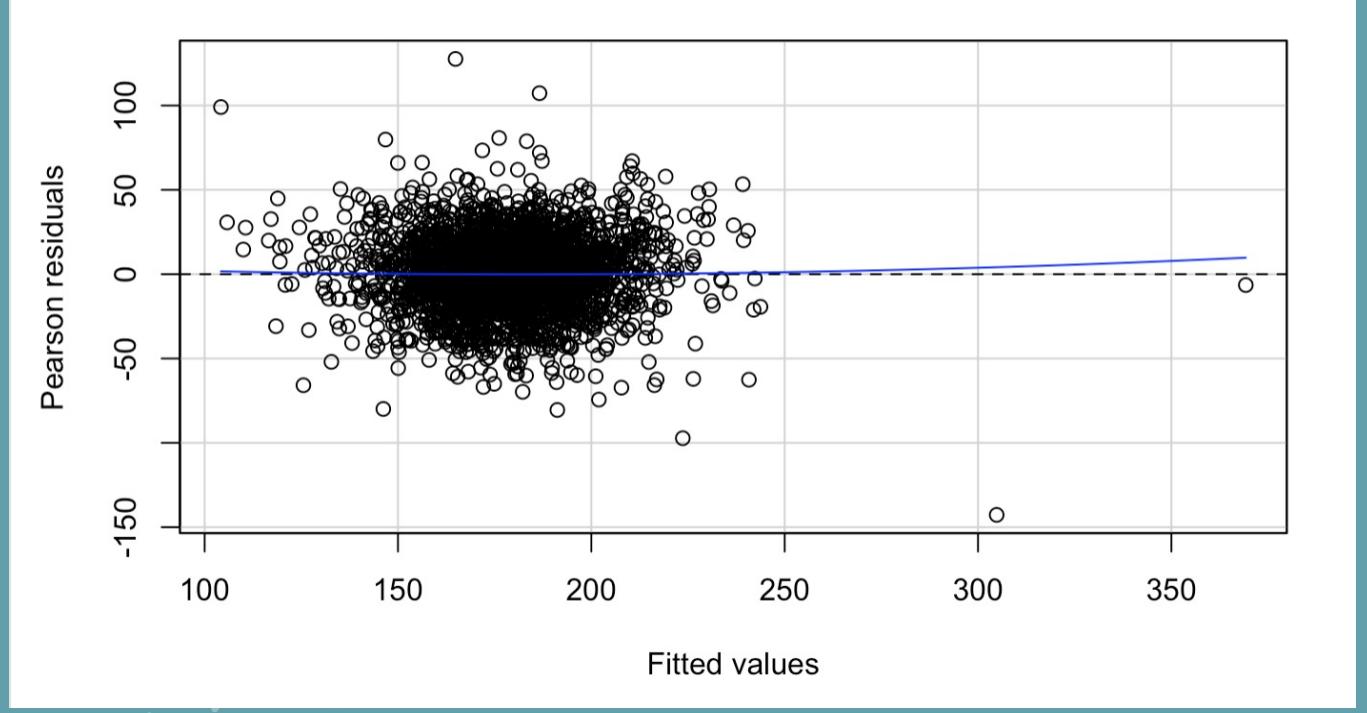


All the histograms show **approximately normal distributions** because of the **bell shaped curve** that is evident and indicates **normality.**

ASSUMPTIONS TESTED

5. HOMOSCEDASCITY: Is there homogeneity of variance?

```{r}
# we use the car package to create a residual plot.
library(car)
residualPlot(cancermodel)
```



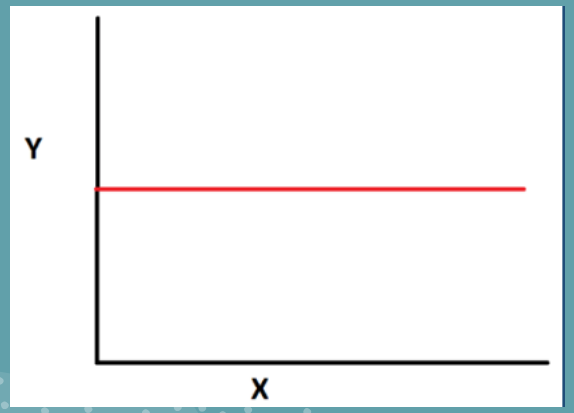
Conclusion: looking at the residual vs fitted plots, we see that the plot is unbiased and homoscedastic as the values are randomly placed on the plot with **no particular trend**. The residual mean is 0, and the standard deviation is constant across the plot. This means that the variance is uniform throughout the range of fitted values. This is also an indication of linearity.

HYPOTHESIS 1

Null hypothesis: Socioeconomic factors do not affect the rates of death due to cancer.

Alternate hypothesis: Socioeconomic factors affect the rates of death due to cancer.

H_0: $\beta_1 = 0$ (is a flat line)
vs.
H_A: $\beta_1 \neq 0$ (is not a flat line)

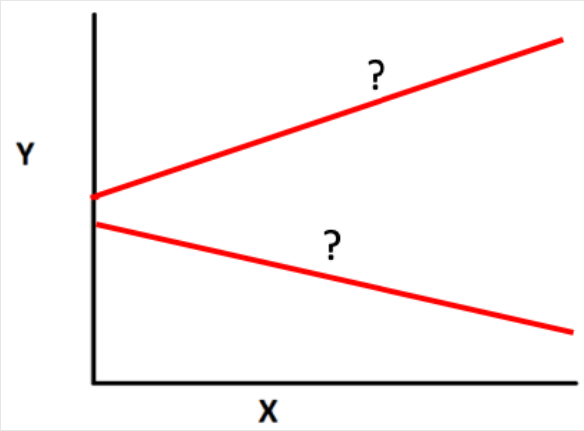


HYPOTHESIS 2

Null hypothesis: Socioeconomic factors increase the rates of death due to cancer.

Alternate hypothesis: Socioeconomic factors decrease the rates of death due to cancer.

H_0: $\beta_1 > 0$ (positive relationship)
vs.
H_A: $\beta_1 < 0$ (negative relationship)



SETTING SIGNIFICANCE LEVEL (alpha)

An alpha ∝ of 0.05 means that 1 out of every 20 times we collect data to run this test, we accept that we will reject the null hypothesis when that's the wrong answer. There are three things we have to consider:

1. **Cost of getting new data:** it would cost a lot to collect such a large amount of data from so many individuals in different counties and would take a lot of time as well.
2. **The risk of making an incorrect decision based on this test:** Since this test is being done for personal research purposes only, there is no risk in making decisions for this test.
3. **3. Ethical considerations associated with someone else using our results to make their decisions**: This data can be used by others to implement changes in the economy to reduce the mortality rates, therefore it could have ethical implications

We would benefit from increasing the significance level to 0.01 to increase accuracy.

RESULTS (HYPOTHESIS 1):

```r
cancermodel <- lm(death_rate ~ pov_percent + incidence_rate
                  + med_income + pvt_coverage)
summary(cancermodel)
```

```
Call:
lm(formula = death_rate ~ pov_percent + incidence_rate + med_income +
    pvt_coverage)

Residuals:
     Min       1Q   Median       3Q      Max
-142.717  -12.509   -0.367   12.610  127.620

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.235e+02  6.580e+00  18.771  < 2e-16 ***
pov_percent     3.714e-01  1.219e-01   3.046  0.00234 **
incidence_rate  2.398e-01  7.171e-03  33.440  < 2e-16 ***
med_income     -4.656e-04  5.308e-05  -8.771  < 2e-16 ***
pvt_coverage   -5.704e-01  6.616e-02  -8.621  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.15 on 3042 degrees of freedom
Multiple R-squared:  0.4198,    Adjusted R-squared:  0.419
F-statistic: 550.2 on 4 and 3042 DF,  p-value: < 2.2e-16
```

Very small p-values

Small p-values for each variable.
Overall p-value = 2.2E-16 < 0.01 (∝)

Since p-value is less than alpha,
we can **REJECT THE NULL HYPOTHESIS.**

CONCLUSION: socioeconomic factors HAVE an effect on cancer mortality rates.

We can now explore Hypothesis 2:

```r
coefficients(cancermodel)
```

```
  (Intercept)    pov_percent incidence_rate     med_income   pvt_coverage
123.517976351    0.371376392    0.239802093   -0.000465566   -0.570379673
```

Death Rate ~ Poverty Percentage

```r
plot(x = pov_percent, y = death_rate, col = "black",
     xlab = "poverty percentage", ylab = "cancer death rate")
mod_pov <- lm(death_rate ~ pov_percent, data = cancer)
abline(mod_pov, col = "lightblue", lwd = 2)
```

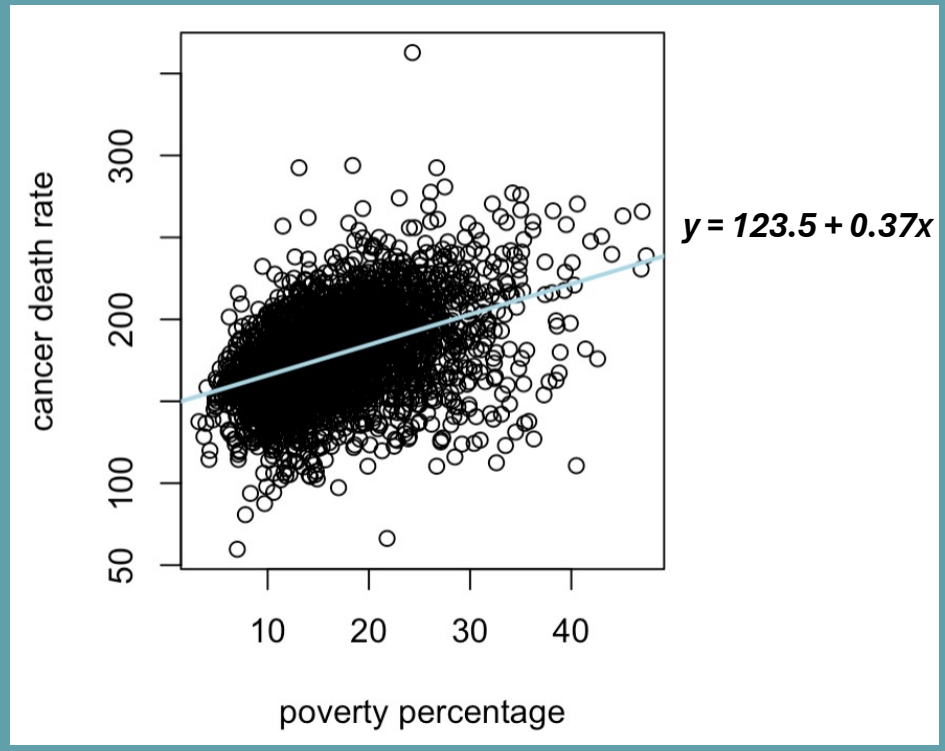


Fig 1. Scatterplot showing the positive relationship between cancer death rate and poverty percentage. As poverty increases, cancer death rate increases. Blue line depicts the line of best fit.

Death Rate ~Incidence Rate

```r
plot(x = incidence_rate, y = death_rate, col = "black",
     xlab = "incidence rate of cancer", ylab = "cancer death rate")
mod_inc <- lm(death_rate ~ incidence_rate, data = cancer)
abline(mod_inc, col = "cyan", lwd = 2)
```

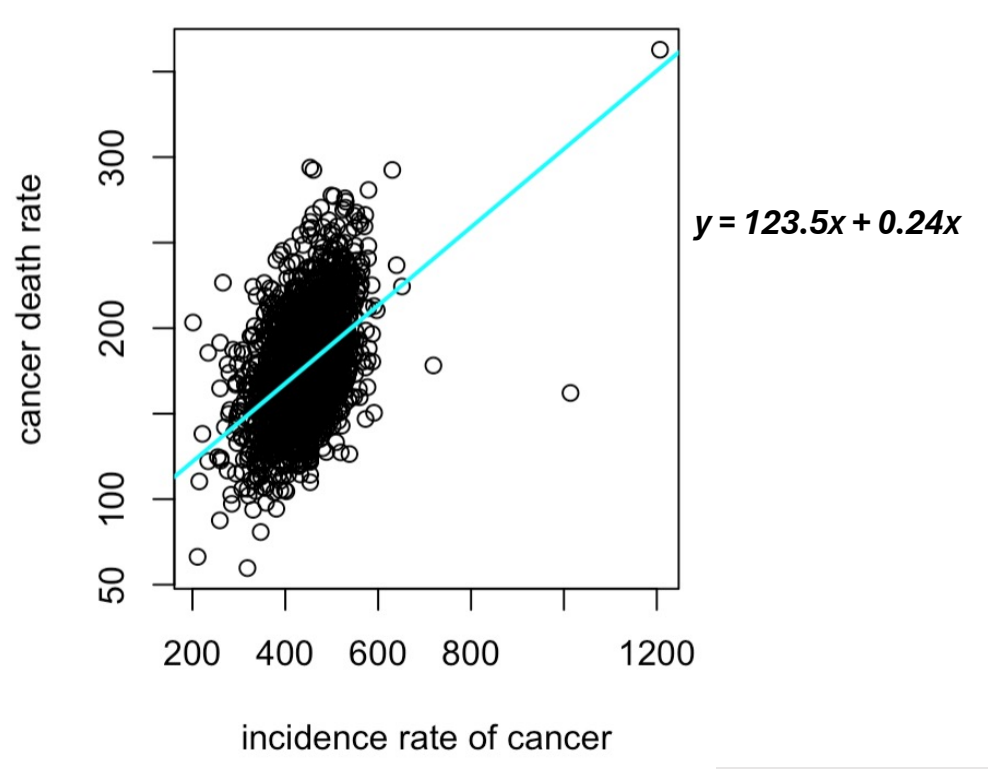


Fig 2. Scatterplot showing the positive relationship between cancer death rate and cancer incidence rate. As the incidence of cancer, deaths due to cancer increases. Aqua blue line depicts the line of best fit.

Death Rate ~ Median Income

```r
plot(x = med_income, y = death_rate, col = "black",
     xlab = "median income", ylab = "cancer death rate")
mod_med_inc <- lm(death_rate ~ med_income, data = cancer)
abline(mod_med_inc, col = "blue", lwd = 3)
```

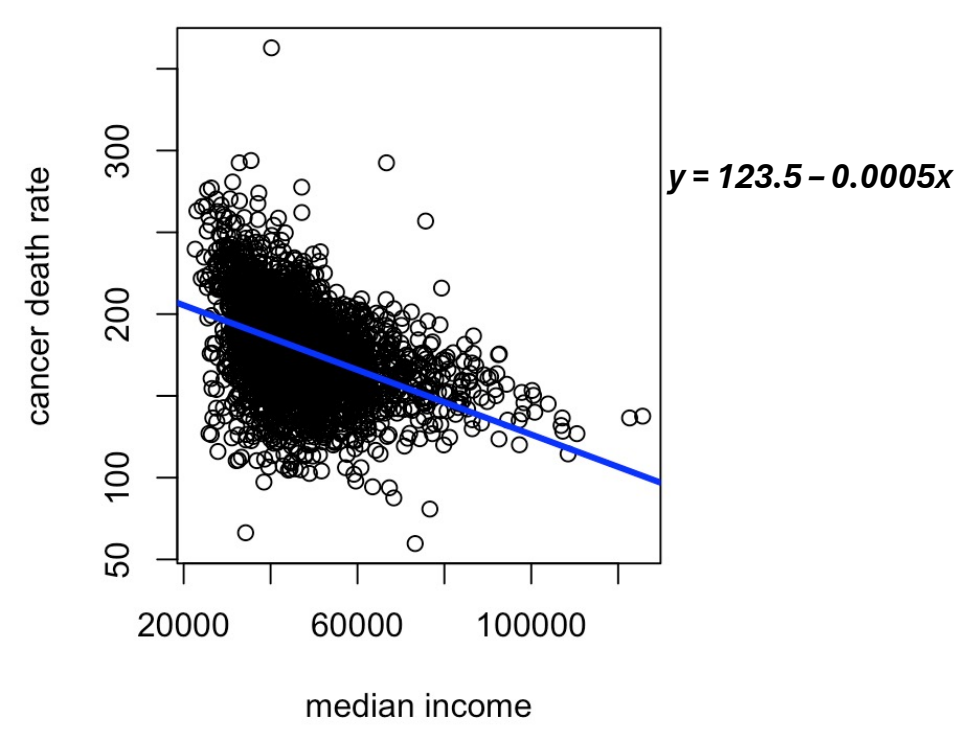


Fig 3. Scatterplot showing the negative relationship between cancer death rate and median income. As median income increases, cancer death rate decreases. Blue line depicts the line of best fit.

Death Rate ~ Private Health Insurance

```r
plot(x = pvt_coverage, y = death_rate, col = "black",
     xlab = "private health insurance", ylab = "cancer death rate")
mod_pvt <- lm(death_rate ~ pvt_coverage, data = cancer)
abline(mod_pvt, col = "purple", lwd = 3)
```

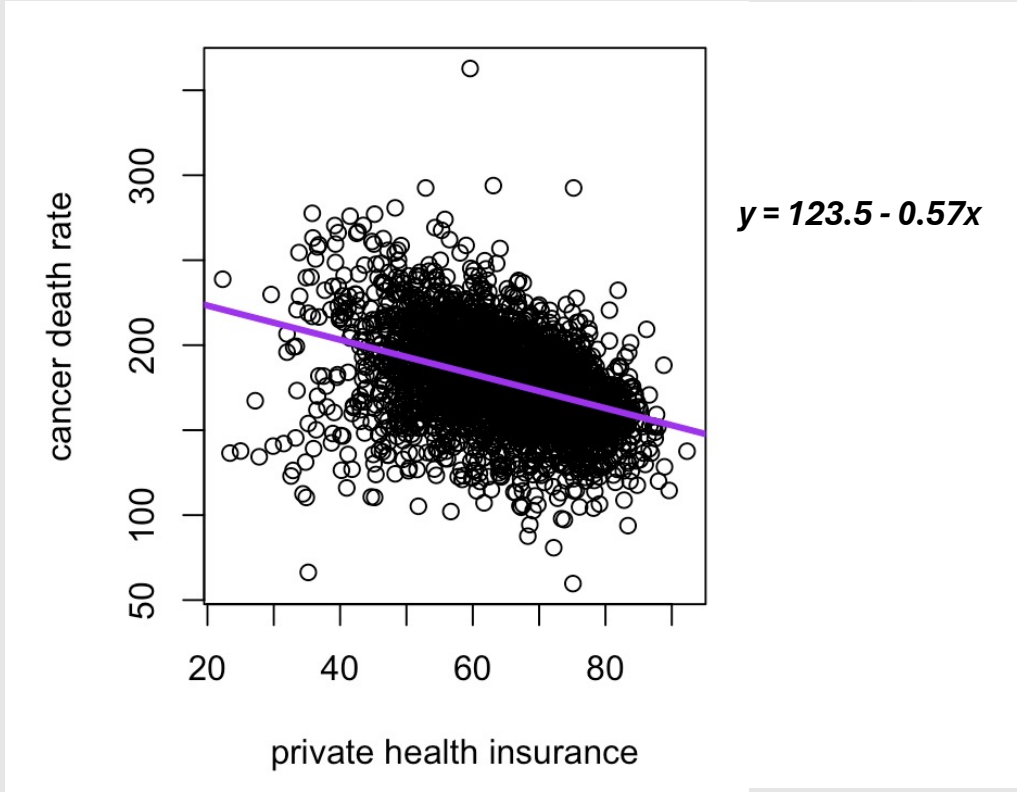


Fig 4. Scatterplot showing the negative relationship between cancer death rate and access private health insurance. As access to private insurance increases, deaths due to cancer decreases. purple line depicts the line of best fit

DISCUSSION

The **estimates** represent the **correlation** of the factor variables to the target variable, with **positive values** for **poverty percent and incidence rate** and a **negative values** for **median income and access to private health insurance.**
The **adjusted R square value** ($R^2 = 0.42$) indicates that there is a **correlation of 42%** between the factor variables and the target variable, and 42% of the variation is explained by the model.
The **slope** of the lines were **statistically significant** (i.e., not a flat horizontal line; p-value < 0.01)

INTERPRETATION OF RESULTS

Fig 1: Positive slope – poverty ⬆ ~ death rate ⬆
Reasons? Lack of access to healthcare, more exposure to carcinogens, low-income jobs.

Fig 2: Positive slope – incidence rate ⬆~ death rate ⬆
Reasons? exposure to carcinogens/habits, increased chances of getting cancer and higher chance of death.

Fig 3: Negative slope – median income ⬆~ death rate ⬇
Reasons? higher income, more access to treatments, higher chance of survival, lower chance of death.

Fig 4: Negative slope – private coverage ⬆ ~ death rate ⬇
Reasons? private coverage ensures that an employee of a private company will get access to treatments that are financially covered by the company.

CONCLUSION

A **multiple linear regression** was conducted to the socioeconomic factors – **poverty percent, cancer incidence rate, median income and access to private health insurance**, affected the **cancer mortality rates.**

We were able to reject the null in our first hypothesis and concluded that the socioeconomic factors of interest do affect the cancer death rate **(**$R^2 = 0.42$, $F_{4, 3042} = 550.2$, $p < 0.01$)

In our second hypothesis, we found significant **positive relationships** between cancer death rate and the factors **poverty percentage** (estimate = +0.37) and **incidence rate** (estimate = +0.24) **.**
We found significant **negative relationships** between cancer death rate and the factors **median income** (estimate = -0.0005) and **private health insurance** (estimate = -0.57) **.**

The extent to which each socioeconomic factor will affect the death rates of cancer in various counties can be estimated using the following slope equations:

- **cancer death rate (%) = 123.5 + 0.37 * Poverty Percent (%)**
- **cancer death rate (%) = 123.5 + 0.24 * Incidence Rate (%)**
- **cancer death rate (%) = 123.5 – 0.0005 * Median Income (USD per annum)**
- **cancer death rate (%) = 123.5 – 0.57 * Access to private health insurance (%)**

THANKYOU FOR LISTENING!

any questions?