CANCER PREVALENCE ON ECONOMICALLY DEPRIVED COUNTIES IN THE UNITED STATES

alex adenuga

2023-01-19

Abstract

In this report, we explored the United States Cancer Data that has 3047 counties across the continental states (excluding Alaska and Hawaii) to determine the best fit Multiple regression model that tells us if the economically deprived regions are home to people who suffer from incidence of cancer and if less clinical trials are performed in these poorer regions.

Results: The model built around the two questions stated above shows that approximately 40% of the variation was explained by the explanatory variables for death rate and the model is significant at 99%. Furthermore, even though the model for Study Per Cap is significant at 99%, it was noted that the variation explained by the explanatory variable is too small (1.7%) and the Residual Standard Error is quite large (525.1). Hence, any prediction done with the model is considered unreliable.

Conclusion: It was concluded that the regions that are economically deprived suffer more cancer incidents and conclusion could not be reached on clinical trials due to insufficient information as seen in the model.

Introduction

Cancer is a blanket term for a family of diseases that affect mammals, where uncontrolled cellular growth occurs in the body. The growth that happens due to cancer can lead to death in some case; especially if it is not detected early or if proper treatment is not in place to stop its growth by killing its cells.

Objectives

The objectives of this report are to:

(i) Determine if the economically deprived regions are home to people who suffer from incidence of cancer.

(ii) Determine if less clinical trial are performed in the poorer regions.

Methodology

A multiple regression Model was used to explore this data.

Arbitrarily written as:

$$y = \beta_0 + \beta_1 x_1 + ...\beta_n x_n + \epsilon$$

Analysis and Results

Loading data

```
cancer<-read.csv("cancer_reg.csv")
```

Removing the variables that are not needed

```
cancer1<-cancer[, -c(8, 9, 13, 16,17,18,19,20,21,22,23
                     ,24,25,26,27,28,29, 30, 31, 32,
                     33, 34)]
```

Checking for correlation with a plot

```
plot(cancer1)
```

Confirming correlation check for some selected variables

```
cor(cancer1)[c(1, 2, 5, 6, 7, 9, 10, 12), c(1, 2, 5, 6, 7, 9,
                                            10, 12)]
```

```
##                 avgAnnCount avgDeathsPerYear   medIncome  popEst2015
## avgAnnCount       1.0000000       0.93940778  0.26914468  0.92689354
## avgDeathsPerYear  0.9394078       1.00000000  0.22320676  0.97763406
## medIncome         0.2691447       0.22320676  1.00000000  0.23552286
## popEst2015        0.9268935       0.97763406  0.23552286  1.00000000
## povertyPercent   -0.1356939      -0.06691794 -0.78896524 -0.06529915
## MedianAgeMale    -0.1249686      -0.14848720 -0.09166264 -0.17660764
## MedianAgeFemale  -0.1228441      -0.14406921 -0.15327840 -0.17793232
## PercentMarried   -0.1061077      -0.18102911  0.35512286 -0.16046328
##                 povertyPercent MedianAgeMale MedianAgeFemale PercentMarried
## avgAnnCount         -0.13569391   -0.12496861      -0.1228441     -0.1061077
## avgDeathsPerYear    -0.06691794   -0.14848720      -0.1440692     -0.1810291
## medIncome           -0.78896524   -0.09166264      -0.1532784      0.3551229
## popEst2015          -0.06529915   -0.17660764      -0.1779323     -0.1604633
## povertyPercent       1.00000000   -0.21400105      -0.1481635     -0.6428569
## MedianAgeMale       -0.21400105    1.00000000       0.9336961      0.4499862
## MedianAgeFemale     -0.14816354    0.93369610       1.0000000      0.3752080
## PercentMarried      -0.64285687    0.44998617       0.3752080      1.0000000
```

NOTE: It appears that popEst2015 is highly correlated with both avgAnncount and AvgDeathperyear.Additionally, MedianAgemale is highly correlated with medianAgefemale. Percentmarried is highly correlated with povertypercent. Finally, it appears that medIncome is highly correlated with povertyPercent.

The Model for Objective(i)

Note: In the model below; explanatory variables that are linearly correlated with another (as shown above) and pose adverse effect on explaining the response variable were removed from the model. In addition, an interaction was observed between PercentMarried & avgAnnCount; this was added to the model to improve the extent to which the response variable was explained by the explanatory variables.

```r
mod1<-lm(TARGET_deathRate~.- MedianAge -  MedianAgeMale-
           MedianAgeFemale - popEst2015 - avgDeathsPerYear
         - medIncome + PercentMarried:avgAnnCount,
         data = cancer1)

summary(mod1)
```

```
##
## Call:
## lm(formula = TARGET_deathRate ~ . - MedianAge - MedianAgeMale -
##     MedianAgeFemale - popEst2015 - avgDeathsPerYear - medIncome +
##     PercentMarried:avgAnnCount, data = cancer1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -149.595  -12.913    0.095   13.195  117.713
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  2.247e+01  7.090e+00   3.170 0.001542 **
## avgAnnCount                  7.315e-03  2.117e-03   3.456 0.000556 ***
## incidenceRate                2.369e-01  7.292e-03  32.493  < 2e-16 ***
## povertyPercent               1.935e+00  8.279e-02  23.377  < 2e-16 ***
## AvgHouseholdSize             6.969e-02  9.213e-01   0.076 0.939705
## PercentMarried               3.664e-01  7.938e-02   4.616 4.08e-06 ***
## avgAnnCount:PercentMarried -2.062e-04  4.583e-05  -4.499 7.07e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.45 on 3040 degrees of freedom
## Multiple R-squared:  0.4039, Adjusted R-squared:  0.4027
## F-statistic: 343.3 on 6 and 3040 DF,  p-value: < 2.2e-16
```

Discussions on objective (i): if the economically deprived regions are home to people who suffer from incidence of cancer?

Related Model:

$$

TargetDeathRate = 22.47 + 0.007AvgAnnCount + 0.237incidenceRate\+ 1.935povertPercent + 0.070AvgHouseholdSize + 0.366percentMarried - \0.0002avgAnncount : Percentmarried

$$

The model for Target Death Rate above shows that The percentage of explained variation is 40.27% and the entire model is significant at 99% with RSE of 21.45.

Additionally, the model shows that 10% increase in the percentage of those in poverty is associated with approximately 19 units increase in death rate from cancer when all other variables are held constant.

This concludes that the economically deprived regions are home to people who suffer from incidence of cancer as the death rate from cancer almost double in outcome due to a unit increase in poverty percent in these regions.

The Analysis and model for objective (ii)

```
cancer2<-cancer[, -c(3, 9, 13, 16,17,18,19,20,21,22,
                     23,24,25,26,27,28,29, 30, 31, 32, 33, 34)]
```

The model for Clinical Trials

```
mod2<-lm(studyPerCap~.- MedianAge -  MedianAgeMale-
           MedianAgeFemale - popEst2015 - avgDeathsPerYear
         - medIncome + PercentMarried:avgAnnCount,
         data = cancer2)
summary(mod2)
```

```
##
## Call:
## lm(formula = studyPerCap ~ . - MedianAge - MedianAgeMale - MedianAgeFemale -
##     popEst2015 - avgDeathsPerYear - medIncome + PercentMarried:avgAnnCount,
##     data = cancer2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -737.4 -160.9 -115.7  -45.1 9452.8
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  4.000e+02  1.736e+02   2.305 0.021237 *
## avgAnnCount                  4.828e-02  5.181e-02   0.932 0.351522
## incidenceRate                6.157e-01  1.785e-01   3.449 0.000570 ***
## povertyPercent              -9.301e+00  2.027e+00  -4.589 4.63e-06 ***
## AvgHouseholdSize            -1.202e+00  2.255e+01  -0.053 0.957495
## PercentMarried              -7.164e+00  1.943e+00  -3.686 0.000231 ***
## avgAnnCount:PercentMarried -6.287e-04  1.122e-03  -0.560 0.575320
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 525.1 on 3040 degrees of freedom
## Multiple R-squared:  0.01912,    Adjusted R-squared:  0.01718
## F-statistic: 9.876 on 6 and 3040 DF,  p-value: 8.18e-11
```

Discussions on objective (ii): if less clinical trial are performed in the poorer regions?

The model for StudyPerCap above shows that The percentage of explained variation is 1.72% even though the entire model is significant at 99%, the RSE is also very large at 525.1 and the extent to which the explanatory variable explained the variation in the response variable is too low. i.e, the Adjusted R squared is too low for any suggested prediction to be reliable.

Conclusion

The models show that those who live in the economically deprived regions predictably suffer more death rates due to cancer and the model to explain clinical trials has a very low variation of the response variable explained; hence, it is not considered reliable to make any form of prediction.