HW8

Simon Hans Edasi

2023-03-03

Multiple Regression: Urban Mortality Rates. In this lab, you will use data from file mortality.csv that is available in Canvas (Files/Lab Datasets). The data contained in this file are from a study in which researchers were studying factors associated with mortality rates in 60 urban areas in the United States in the 1960s; we will be using a subset of the original dataset.

The file contains 7 columns of data collected from each of the 60 urban areas. The columns include:

Size: the mean household size

School: the median number of school years completed by those over 22

Density: Population per square mile

Nonwhite: Percentage of the non-white population

NO: Measure of the levels of nitric oxides as a pollutant

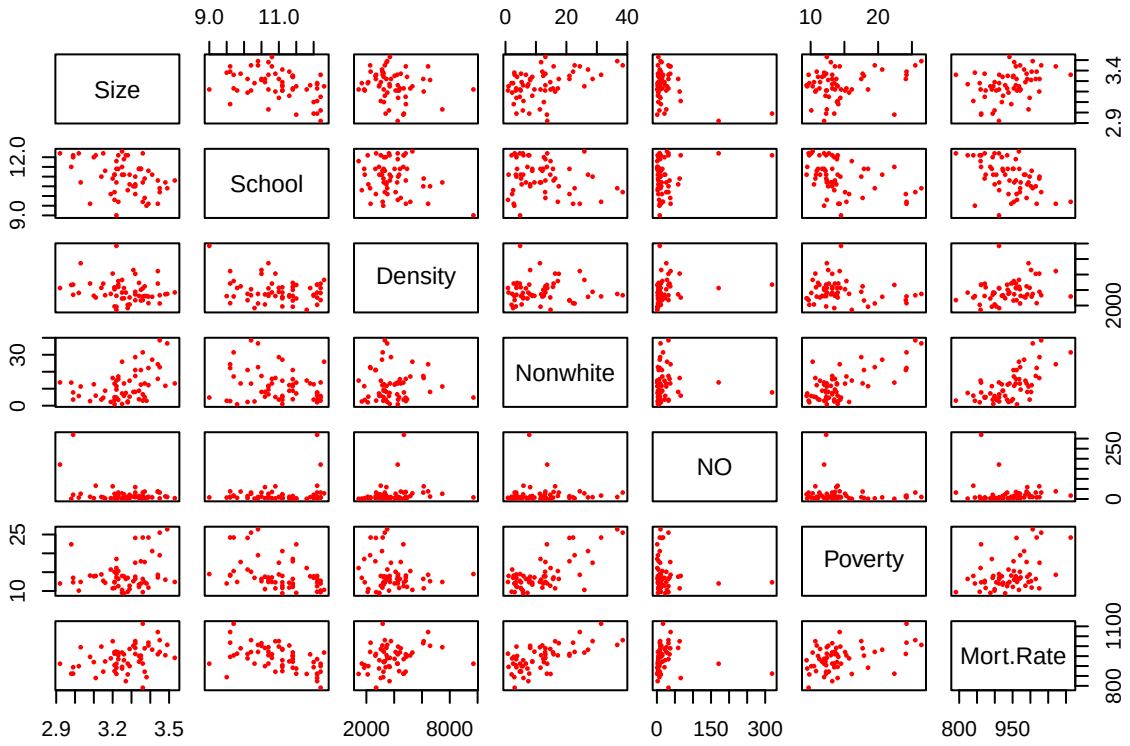
Poverty: Percentage of families with an income below the poverty level

Mort.Rate: the mortality rate per 100,000 residents

We will be using multiple regression to measure the effect of the predictor variables Size, School, Density, Nonwhite, NO, and Poverty on the response variable, Mort.Rate.

(1) Before conducting a regression analysis, let's explore the dataset by a plotting pair-wise scatterplot using the plot() command (recall your code from Lab 7). Add a color of your choice and paste your plot below. (6 points)

```
mortality <- read.csv('mortality.csv')
attach(mortality)
plot(mortality, pch = 16, cex = 0.5, col = 'red')
```



(2) In your pair-wise scatterplot, which, if any, predictor variables appear to be correlated with Mort.Rate? Which, if any, predictor variables do not appear to be correlated with Mort.Rate? (12 points)

Mortality rate appears to be correlated with Poverty, Nonwhite, and School.

(3) Multiple linear regression assumes that the response variable is normality distributed. Plot a histogram of the response variable, Mort.Rate. Include a title and a color of your choice, and paste your histogram below. (6 points)

```
hist(Mort.Rate, col = 'light blue', main = 'Histogram of Mortality Rate')
```

(4) What is the mean and median of Mort.Rate? Based on a visual assessment of your histogram in (3), and your estimates of the mean and median, do you conclude that Mort.Rate is normally distributed? Why or why not? (9 points)

```
mean(Mort.Rate)
```

```
## [1] 940.3584
```

```
median(Mort.Rate)
```

```
## [1] 943.683
```

The mean and median are about the center of the semi-bell curve. There is no heavy skew and kurtosis looks fine. I think this is more-or-less normally distributed.

```
shapiro.test(Mort.Rate)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Mort.Rate
## W = 0.99375, p-value = 0.99
```

(5) Regardless of your conclusion in (4), let's assume Mort.Rate is normally distributed, and let's use multiple regression to determine which, if any, of the predictor variables can be used to statistically predict Mort.Rate. Run a multiple regression and include all predictor variables from the dataset when using Mort.Rate as the response variable. Paste your code and output below. (6 points)

```
summary(model <- lm(Mort.Rate~Poverty+NO+Nonwhite+Density+School+Size))
```

```
##
## Call:
## lm(formula = Mort.Rate ~ Poverty + NO + Nonwhite + Density +
##     School + Size)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -101.151  -16.883   -1.429   22.524   97.217
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.434e+03  2.563e+02   5.596 7.90e-07 ***
## Poverty     -4.693e+00  2.198e+00  -2.135 0.037429 *
## NO          -8.994e-02  1.260e-01  -0.714 0.478405
## Nonwhite     5.624e+00  1.007e+00   5.584 8.25e-07 ***
## Density      4.294e-03  4.384e-03   0.980 0.331775
## School      -3.418e+01  8.584e+00  -3.982 0.000209 ***
## Size        -4.061e+01  5.413e+01  -0.750 0.456512
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.02 on 53 degrees of freedom
## Multiple R-squared:  0.6282, Adjusted R-squared:  0.5861
## F-statistic: 14.92 on 6 and 53 DF,  p-value: 6.631e-10
```

(6) Test the null hypothesis, Ho: estimate = 0, for each predictor variable using alpha=0.05, and indicate your statistical conclusion for each (12 points)

Poverty p_value = 0.04 - reject H_o
NO p_value = 0.48 - Fail to reject H_o
Nonwhite p_value = $8*10^{-7}$ - reject H_o
Density p_value = 0.33 - Fail to reject H_o
School p_value = 0.0021 - reject H_o
Size p_value = 0.46 - Fail to reject H_o

(7) Conduct a correlation test between Mort.Rate and each significant predictor variable (when using alpha = 0.05) from your output in (5), and list the correlation coefficients for each below. (6 points)

```
cor.test(Mort.Rate, Poverty)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Mort.Rate and Poverty
## t = 3.4283, df = 58, p-value = 0.001124
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1747805 0.6016955
## sample estimates:
##       cor
## 0.4104873
```

Correlation coefficient between mortality rate and poverty is 0.41, and confidence intervals do not contain 0.

```
cor.test(Mort.Rate, NO)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Mort.Rate and NO
```

```
## t = -0.61465, df = 58, p-value = 0.5412
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3276772  0.1770962
## sample estimates:
##         cor
## -0.08044605
```

Correlation coefficient between mortality rate and Nitrous Oxide leakage is -0.080 and confidence intervals contain 0.

```
cor.test(Mort.Rate, Nonwhite)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Mort.Rate and Nonwhite
## t = 6.4071, df = 58, p-value = 2.88e-08
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4660302 0.7715694
## sample estimates:
##      cor
## 0.643773
```

Correlation coefficient between mortality rate and nonwhite is 0.64 and confidence intervals do not contain 0.

```
cor.test(Mort.Rate, Density)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Mort.Rate and Density
## t = 2.0972, df = 58, p-value = 0.04034
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.01240982 0.48661662
## sample estimates:
##       cor
## 0.2654979
```

Correlation coefficient between mortality rate and density is 0.27 and confidence intervals do not contain 0.

```
cor.test(Mort.Rate, School)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Mort.Rate and School
## t = -4.5272, df = 58, p-value = 3.022e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.6770590 -0.2953857
## sample estimates:
##        cor
## -0.5109837
```

Correlation coefficient between mortality rate and school is -.51 and confidence intervals do not contain 0.

```
cor.test(Mort.Rate, Size)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Mort.Rate and Size
## t = 2.9136, df = 58, p-value = 0.005068
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1137070 0.5603944
## sample estimates:
##       cor
## 0.3573149
```

Correlation coefficient between mortality rate and size is 0.38 and confidence intervals do not contain 0.

(8) Interpret the correlation coefficients from (7); how does each significant predictor variable relate to Mort.Rate? (6 points)

Positive values of correlation coefficients mean positive correlation of variables, and distance from 0 represents how much the variables co-vary. From highest to lowest positive correlation with mortality rate we have nonwhite, poverty, size, and density. School is negatively correlated with mortality rate as is nitrous oxide leakage, however NO is very weakly correlated and the confidence intervals include zero, which leads to a possibility that mortality rate and NO is not correlated at all.

(9) Based on your output from (5), what is the final regression equation? (6 points)

mortality rate = (-4.69 * poverty) + (-899.4 * NO) + (5.62 * nonwhite) + (439.4 * density) + (-34.18 * school) + (-40.61 * size) + 1434

(10) Using your output from your multiple regression model in (5), how much of the variation in Mort.Rate is explained by the predictor variables? (6 points)

From the adjusted $R^2 = 0.5861$, we can explain about 59% of mortality rate variation with this regression model.