Extra Assignment

Evan Hodges

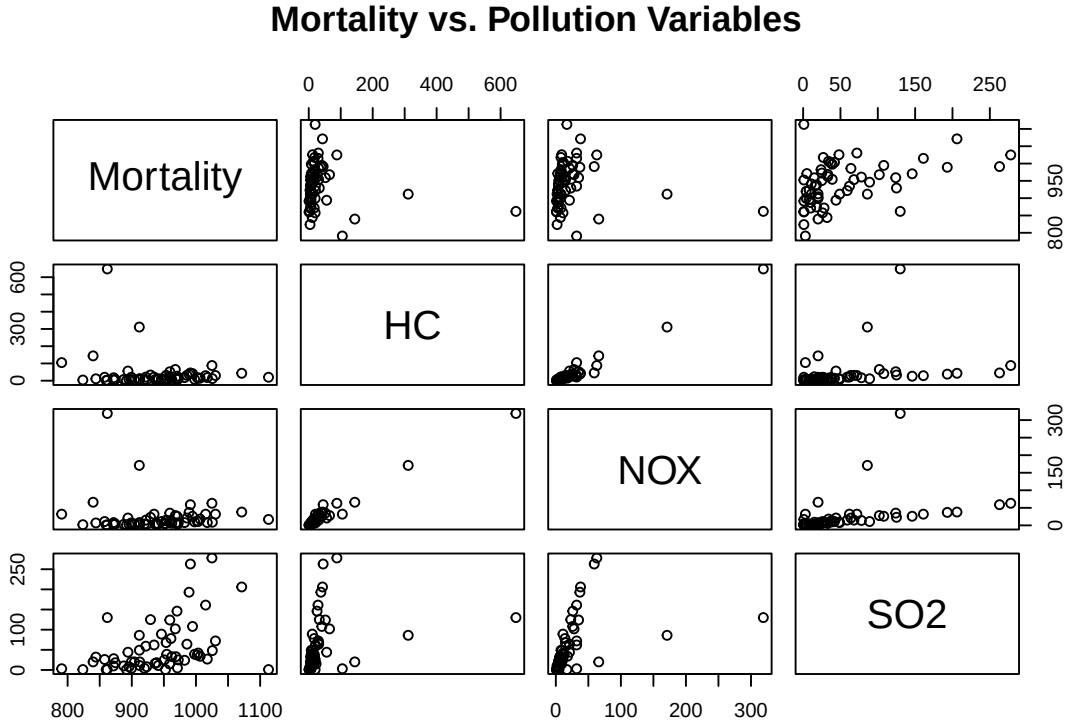
2025-04-11

1. Does pollution kill people? Total age-adjusted mortality from all causes, in deaths per 100,000 population, is the response variable. The 15 variables for each of 60 cities are (1) mean annual precipitation (in inches); (2) percent relative humidity (annual average at 1 P.M.); (3) mean January temperature(in degrees Fahrenheit); (4) mean July temperature (in degrees Fahrenheit); (5) percentage of the population aged 65 years or over; (6) population per household; (7) median number of school years completed by persons of age 25 years or more; (8) percentage of the housing that is sound with all facilities; (9) population density (in persons per square mile of urbanized area); (10) percentage of 1960 population that is nonwhite; (11) percentage of employment in white-collar occupations;(12) percentage of households with annual income under $3,000 in 1960; (13) relative pollution potential of hydrocarbons (HC); (14) relative pollution potential of oxides of nitrogen (NOX); and (15) relative pollution potential of sulfur dioxide (SO2). It is desired to determine whether the pollution variables (13, 14, and 15) are associated with mortality.

```
mort <- read.csv("mort.csv")
```

   a. Obtain a pairwise scatter plot to explore relationship between mortality and the pollution variables. Comment on your observations.

```
pollution_vars <- mort[, c("Mortality", "HC", "NOX", "SO2")]
pairs(pollution_vars, main = "Mortality vs. Pollution Variables")
```



```r
# From the scatter-plot matrix, you can see a general upward trend between Mortality
# and each of the three pollution measures (HC, NOX, and SO2). Higher pollution
# levels tend to coincide with higher mortality rates, although the strength of that
# relationship may vary by pollutant. There also appears to be some clustering in the
# data: a group of cities with relatively lower pollution and mortality, and another
# set with higher values of both. Additionally, the pollution variables themselves
# show positive associations with each other, suggesting that cities with high levels
# of one pollutant tend to have high levels of the others as well.
```

b. With mortality as the response, fit a regression involving the weather and socioeconomic variable as explanatory variables (independent variables), then use the R function stepAIC() to perform a stepwise regression to select the important variables. Describe the relationship of the selected variables with mortality.

```r
initial_model <- lm(Mortality ~ Precip + Humidity + JanTemp + JulyTemp + Over65
                    + House + Educ + Sound + Density + NonWhite + WhiteCol
                    + Poor, data = mort)
summary(initial_model)
```

```
##
## Call:
## lm(formula = Mortality ~ Precip + Humidity + JanTemp + JulyTemp +
##     Over65 + House + Educ + Sound + Density + NonWhite + WhiteCol +
##     Poor, data = mort)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -72.677 -19.583  -3.084  20.636  82.627
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.770e+03  4.443e+02   3.984 0.000234 ***
## Precip       1.572e+00  8.250e-01   1.906 0.062842 .
## Humidity    -1.145e-01  1.104e+00  -0.104 0.917840
## JanTemp     -2.166e+00  9.995e-01  -2.167 0.035349 *
## JulyTemp    -3.103e+00  1.859e+00  -1.669 0.101750
## Over65      -4.593e+00  8.267e+00  -0.556 0.581169
## House       -1.033e+02  7.238e+01  -1.428 0.160027
## Educ        -2.089e+01  1.122e+01  -1.861 0.068970 .
## Sound       -3.761e-01  1.814e+00  -0.207 0.836618
## Density      5.325e-03  4.174e-03   1.276 0.208298
## NonWhite     5.741e+00  1.157e+00   4.962 9.58e-06 ***
## WhiteCol    -3.992e-01  1.644e+00  -0.243 0.809197
## Poor        -7.119e-01  3.291e+00  -0.216 0.829669
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.68 on 47 degrees of freedom
## Multiple R-squared:  0.723,  Adjusted R-squared:  0.6523
## F-statistic: 10.22 on 12 and 47 DF,  p-value: 1.829e-09
```

```r
library(MASS)
step_model <- stepAIC(initial_model, direction = "both", trace = FALSE)
summary(step_model)
```

```
##
## Call:
## lm(formula = Mortality ~ Precip + JanTemp + JulyTemp + House +
##     Educ + Density + NonWhite, data = mort)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -74.109 -20.783  -1.205  19.554  81.604
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.525e+03  2.308e+02   6.608 2.09e-08 ***
## Precip       1.276e+00  6.075e-01   2.100  0.04063 *
## JanTemp     -2.123e+00  6.089e-01  -3.487  0.00100 **
## JulyTemp    -2.728e+00  1.278e+00  -2.134  0.03758 *
## House       -7.003e+01  4.852e+01  -1.443  0.15492
## Educ        -2.003e+01  7.112e+00  -2.817  0.00684 **
## Density      5.555e-03  3.567e-03   1.557  0.12543
## NonWhite     5.892e+00  8.061e-01   7.309 1.59e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.07 on 52 degrees of freedom
## Multiple R-squared:  0.7198, Adjusted R-squared:  0.6821
## F-statistic: 19.09 on 7 and 52 DF,  p-value: 2.387e-12
```

```r
# In the final stepwise-selected model, Mortality is explained best by
# Precip, JanTemp, JulyTemp, House, Educ, Density, and NonWhite.
#
# Interpreting the regression coefficients:
# Precipitation (Precip) = higher mortality.
# Warmer winter (JanTemp) and summer (JulyTemp) = lower mortality.
# Areas with higher average housing (House) values = lower mortality.
# Higher levels of education (Educ) = lower mortality.
# More densely populated areas (Density) = slightly higher mortality.
# Higher proportions of nonwhite residents (NonWhite) = higher mortality.
```

c. To the model chosen from stepwise regression, add the three pollution variables (transformed to their logarithms). Using the estimated coefficients, describe the relationship between the pollution variables and mortality.

```r
mort$logHC <- log(mort$HC)
mort$logNOX <- log(mort$NOX)
mort$logSO2 <- log(mort$SO2)

final_model <- update(step_model, . ~ . + logHC + logNOX + logSO2)
summary(final_model)
```

```
##
## Call:
## lm(formula = Mortality ~ Precip + JanTemp + JulyTemp + House +
##     Educ + Density + NonWhite + logHC + logNOX + logSO2, data = mort)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -70.515 -15.745   1.561  20.595  62.677
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.460e+03  2.602e+02   5.610 9.28e-07 ***
## Precip       1.903e+00  6.515e-01   2.921  0.00526 **
## JanTemp     -2.548e+00  7.520e-01  -3.389  0.00139 **
## JulyTemp    -2.370e+00  1.630e+00  -1.454  0.15234
## House       -6.976e+01  4.656e+01  -1.498  0.14052
## Educ        -1.616e+01  7.055e+00  -2.290  0.02638 *
## Density      3.873e-03  3.549e-03   1.091  0.28046
## NonWhite     5.254e+00  9.466e-01   5.550 1.15e-06 ***
## logHC       -2.919e+01  1.458e+01  -2.002  0.05086 .
## logNOX       4.553e+01  1.421e+01   3.205  0.00238 **
## logSO2      -7.793e+00  6.593e+00  -1.182  0.24290
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.07 on 49 degrees of freedom
## Multiple R-squared:  0.7792, Adjusted R-squared:  0.7341
## F-statistic: 17.29 on 10 and 49 DF,  p-value: 7.209e-13
```

```
# logHC suggests that higher levels of hydrocarbons correlate with a slight decrease
# in mortality- though this effect is relatively weak and only marginally significant.
# logNOX indicates that higher NO levels are associated with higher mortality.
# logSO2 implies that there is no clear evidence of an SO2-mortality relationship.
```