Pollution and Mortality in Metropolitan Cities

Does pollution kill people? Can living in a large city put a person at significant risk for their life? We examine 15 different variables from the Standard Metropolitan Statistical Areas, obtained from 1959 to 1961 in order to answer these questions:

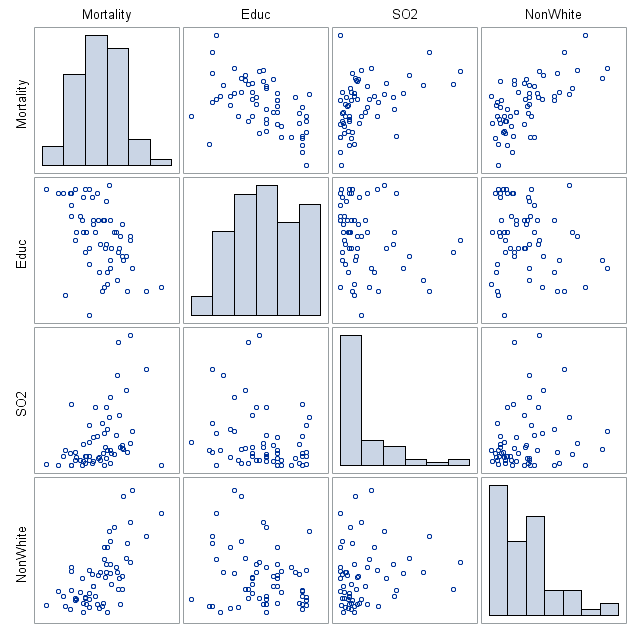
|  |  |
| --- | --- |
| **Variable** | **Additional Remarks** |
| Mean Annual Precipitation (inches) |  |
| Percent Relative Humidity | Annual Average at 1pm |
| Mean January Temperature | Degrees Fahrenheit |
| Mean July Temperature | Degrees Fahrenheit |
| Percentage Population > 65 |  |
| Population Per Household |  |
| Median Number of School Years | Age 25 and Older |
| Percentage of Housing With Facilities |  |
| Population Density | Persons Per Square Mile of Urbanized Area |
| Percentage of Population – Non White |  |
| Percentage of Households Income | With Annual Income < $3000 |
| Relative Pollution Potential (HC) | Hydrocarbons |
| Relative Pollution Potential (NOx) | Oxides of Nitrogen |
| Relative Pollution Potential (SO2) | Sulphur Dioxide |

Initial investigation involved building a correlation matrix for each variable associated with mortality rate. Precipitation (.51), education (-.51), nonwhite (.64), sound facilities (-.43), poor (.41) and sulphur oxide (.43) are moderately to highly correlated with mortality rate.

Based on initial correlations, it’s obvious that pollution is not the only variable associated with mortality rate in urbanized areas for the time period in question. In order to appropriately estimate the effects of variables on mortality rates, we build a multiple regression model to investigate each variable’s statistical significance and impact as part of an overall predictive model for mortality rate.

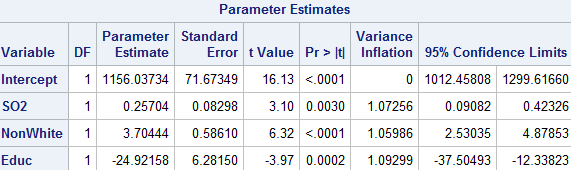
In order to pare down our model, correlations were investigated between potential explanatory variables. Precipitation and sound are moderately correlated with non white (.41, -.43). Further, poor is moderately associated with education (-.41). By including the correlated terms in a regression model, we will likely over fit our model and reduce the impact of other highly correlated variables such as nonwhite and education. Therefore, we will move forward with three variables in our model: education, nonwhite and sulphur oxide.

Scatterplots for education, nonwhite and sulphur oxide are created to visually investigate normality, applicability of a linear model and association between variables further:

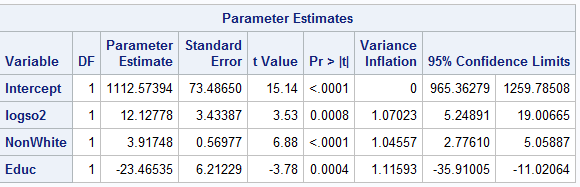


Based on initial investigation, SO2 and Nonwhite are potential candidates for transformation. However, there are obvious linear patterns between the explanatory variables and mortality rate. We will move forward with the base model and work our way up to an optimal multiple regression model.

Residual analysis on the multiple regression model of {Mortality | SO2, NonWhite, Educ} shows abnormal residuals for SO2 an rsquared of 0.626 with an overall F of 31.33 and overall p <.0001. All variables in the model are statistically significant:

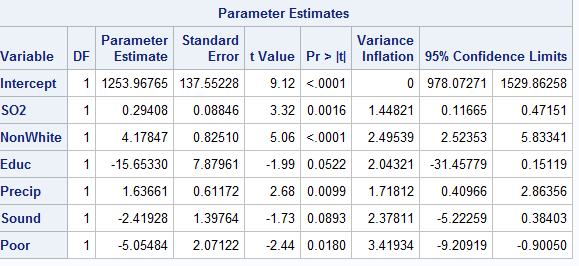


Correcting for the non-normality and non-constant variance in SO2 by log transforming SO2, residuals look randomly scattered and normal for each variable with a larger r-squared value of 0.642. New Orleans represents an outlier with high leverage in our dataset. However, given our sample size, this leverage point is included in the data. The second model’s overall F is higher as well, with a value of 33.52. All three explanatory variables are statistically significant with low levels of cross-correlation:



Of the three pollutants, only sulphur oxide is highly correlated with mortality rate. Other social factors, such as nonwhite and education are also highly correlated with mortality rate.

Just to confirm that our selection of variables is accurate, another regression was ran including precipitation, poor and sound variables. The results are as follows:



Obviously, multiple variables are now insignificant in the model. Further, the variance inflation factor becomes larger than required. Adjusted r-squared only rises by .02 ; from 0.62 to 0.64. An additional three variables for .02 points in adjusted r-squared is too expensive. Along with insignificance of one third of the variables, we are comfortable saying our parsimonious model is more suitable for analysis.

Our final model is:

y’{Mortality Rate | SO2, NonWhite, Educ} = 1112.57394 + 12.12778(ln(SO2)) + 3.91748(NonWhite) – 23.46535(Educ).

All other variables held equal, a doubling of SO2 results in a change of 12.12778(ln(2))) = 8.4 unit increase in mortality rate and a one unit increase in nonwhite results in a rise of 3.92 points in mortality rate. Education has a negative relationship with mortality rate: for each education level completed, mortality rate decreases by 23.47 points.

Inferences from this model cannot be taken as causal as all data used is observational. Further, the time period and setting (urban cities) do not allow inferences outside of the 1959-61 and urban setting. That being said, from 1959 to 1961, education, sulphur oxide potential and nonwhite ethnicities were associated with higher mortality rates in urban settings.

SAS CODE

**PROC** **IMPORT** OUT = Pollution

DATAFILE = "\\Client\C$\Users\patrickcorynichols\Desktop\ex1217.csv"

DBMS = CSV REPLACE;

GETNAMES = YES;

DATAROW = **2**;

**RUN**;

**data** pollution2;

SET pollution;

logso2 = log(so2);

lognonwhite = log(nonwhite);

**RUN**;

**PROC** **CORR** data = pollution;

**RUN**;

**PROC** **sgscatter** data = pollution2;

matrix mortality educ so2 nonwhite / diagonal = (histogram);

**RUN**;

**PROC** **sgscatter** data = pollution2;

matrix mortality educ logso2 lognonwhite / diagonal = (histogram);

**RUN**;

**PROC** **REG** data = pollution2;

MODEL mortality = so2 nonwhite educ precip sound poor / VIF CLB;

**RUN**;

**PROC** **REG** data = pollution2;

MODEL mortality = logso2 nonwhite educ / VIF CLB CLM CLI;

**RUN**;

**PROC** **REG** data = pollution2;

MODEL mortality = logso2 nonwhite / VIF CLB;

**RUN**;

**PROC** **REG** data = pollution2;

MODEL mortality = logso2 lognonwhite educ / VIF CLB;

**RUN**;