Model Selection and Evaluation Worksheet

David Gerard

2018-12-07

Pollution and Mortality

Does pollution kill people? Data in one early study designed to explore this issue came from five Standard Metropolitan Areas (SMSA) in the United States, obtained for the years 1959-1961. The variables in these data are:

- `CITY`: a character vector indicating the city
- `Mortality`: total age-adjusted mortality from all causes
- `Precip`: mean annual precipitation (inches)
- `Humidity`: percent relative humidity (annual average at 1:00pm)
- `JanTemp`: mean January temperature (degrees F)
- `JulyTemp`: mean July temperature (degrees F)
- `Over65`: percentage of the population aged 65 years or over
- `House`: population per household
- `Educ`: median number of school years completed for persons 25 years or older
- `Sound`: percentage of the housing that is sound with all facilities
- `Density`: population density (in persons per square mile of urbanized area)
- `NonWhite`: percentage of population that is nonwhite
- `WhiteCol`: percentage of employment in white collar occupations
- `Poor`: percentage of households with annual income under $3,000 in 1960
- `HC`: relative pollution potential of hydrocarbons
- `NOX`: relative pollution potential of oxides of nitrogen
- `SO2`: relative pollution potential of sulfur dioxide

The goal of the study is to determine if the pollution variables (`HC`, `NOX`, and `SO2`) are associated with mortality after the other climate and socioeconomic variables are accounted for. (*Note*: These data are spatially dependent, but we'll ignore that for this worksheet)

You can load these data into R using:

```
library(Sleuth3)
data("ex1217")
head(ex1217)
```

1. What is the response variable? What are the explanatory variables? What are the observational units? How many observational units are there?
2. Run an exploratory data analysis. Do you notice any variables that might need to be transformed? Do you notice any outlying observations? Apply any transformations where needed and comment on any potential outlying observations.
3. Fit a tentative rich model with all variables except the pollution variables (and no interactions) and comment on any issues you see with the residuals.
4. Calculate Cook's distance, the studentized residuals, and the leverage values in this rich model. Are there any extreme observations?
5. Use `step()` to run a step-wise procedure to select important variables. Make sure you don't include any of the pollution variables in the model. What variables were included?

6. Reevaluate the residuals (and case-influence statistics) from the final final selected by the step-wise procedure. Are the assumptions of the linear model ok?
7. Now we will include the pollution variables to the variables selected in the previous part. Write out the linear regression model.
8. Fit this model and evaluate the residuals and case-influence statistics.
9. Write out the hypothesis tests for testing if **any** of the pollution variables are associated with mortality.
10. Run the above hypothesis test in R. State your conclusion. Include estimates and 95% confidence intervals for the estimated effects of pollution on mortality.
11. If you have time, try re-running the analysis (step-wise regression and F-test) by excluding any high-leverage cities. Do your results change?