STAT W4201 001, Homework 8

Brian Weinstein (bmw2148)

Apr 6, 2016

Code is attached here and also posted at https://github.com/BrianWeinstein/advanced-data-analysis. Where relevant, code snippets and output are are included in-line.

Problem 1: Ramsey 12.17

It is desired to determine whether the pollution variables (13, 14, and 15) are associated with mortality, after the other climate and socioeconomic variables are accounted for. (Note: These data have problems with influential observations and with lack of independence due to spatial correlation; these problems are ignored for purposes of this exercise.)

(a) *With mortality as the response, use a Cp plot and the BIC to select a good-fitting regression model involving weather and socioeconomic variables as explanatory. To the model with the lowest Cp, add the three pollution variables (transformed to their logarithms) and obtain the p-value from the extra-sum-of-squares F-test due to their addition.*

A matrix of pairwise scatterplots is shown in Figure 1.

Initial investigation of the pairwise scatterplots indicate a couple of things.

- The effect of the JanTemp, JulyTemp, Over65, and House variables on Mortality are all nonlinear. As each of these variables increases (marginally, at least), Mortality first increases and then decreases. Adding quadratic terms for each of these variables will help model this behavior.
- Many of the explanatory variables are highly correlated (e.g., JanTemp and JulyTemp, Educ and Sound, Humidity and WhiteCol, etc.). The inclusion of all of these variables will likely be unnecessary in the final model, but we keep them here so that it's unlikely we miss an important relationship during initial investigation.

Using the *leaps::regsubsets* R function, we first do an exhaustive search over all 2^16 possible models (using the original 12 weather and socioeconomic variables, and the 4 squared terms), recording the Cp and BIC for the best few models of each size. We then ignore the models that include a squared variable but don't include the associated linear variable (as per section 12.6).

A Cp plot is shown in Figure 2 for the remaining models. To reduce clutter, I've only included those models with relatively low Cp statistics.

The model with the lowest Cp statistic is the one that includes Precip, JanTemp, JulyTemp, Educ, Density, and NonWhite, plus an intercept term. For this model, the Cp statistic is 1.922. This model has the 3rd lowest BIC, and we continue with this set of variables for the remainder of the problem.

To the 6-variable model, we add the log-transformed pollution variables and perform an extra-sum-of-squares F-test as shown below.

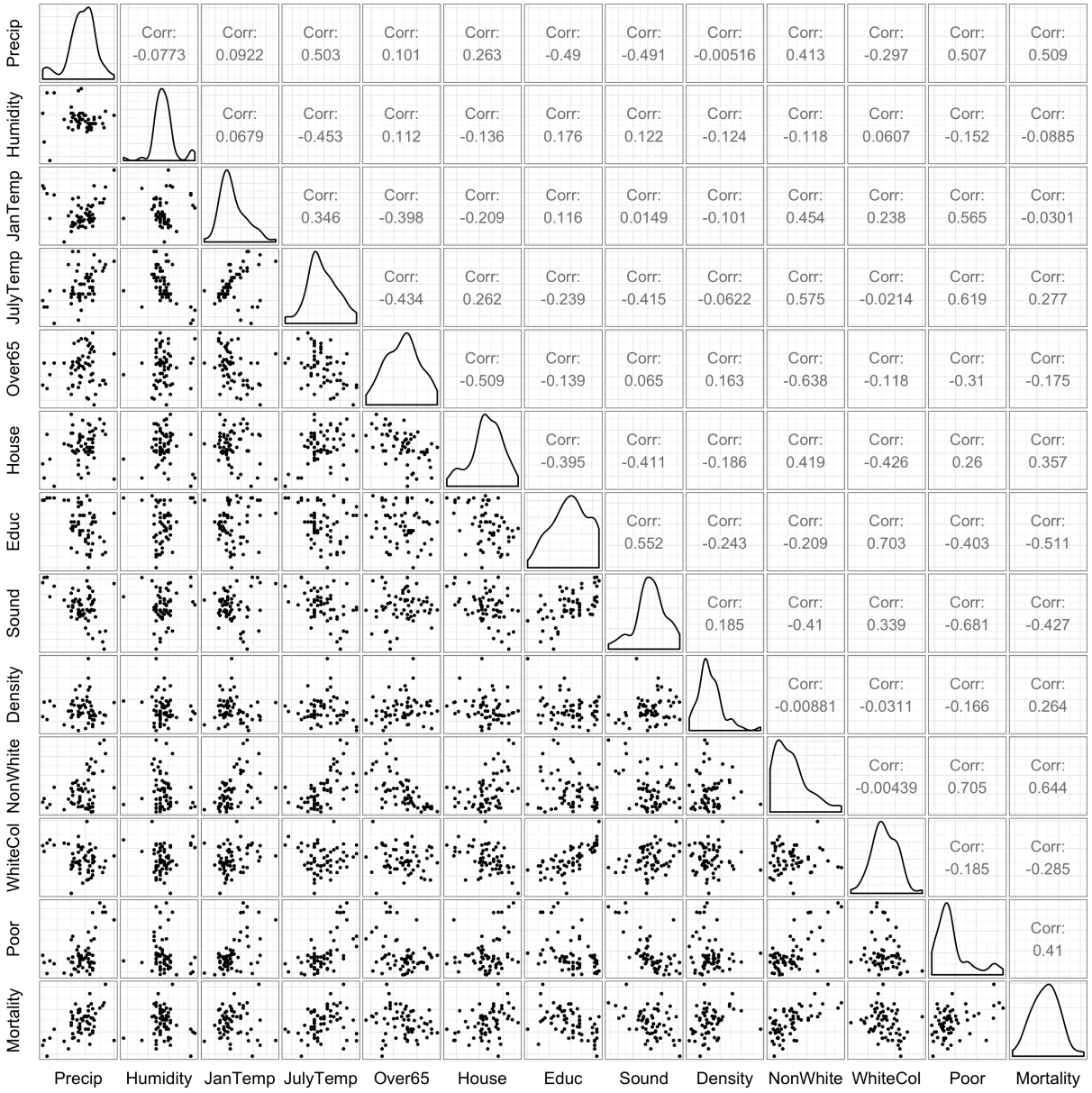

Figure 1: Pairwise scatterplots of the weather and socioeconomic varibles from the "Pollution and Mortality" dataset.

```
> anova(lmBestSubsetPoll, lmBestSubset)
Analysis of Variance Table

Model 1: Mortality ~ Precip + JanTemp + JulyTemp + Educ + Density + NonWhite +
    log(HC) + log(NOX) + log(SO2)
Model 2: Mortality ~ Precip + JanTemp + JulyTemp + Educ + Density + NonWhite
  Res.Df   RSS Df Sum of Sq     F   Pr(>F)
1     50 52712
2     53 66518 -3    -13806 4.365 0.008313 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

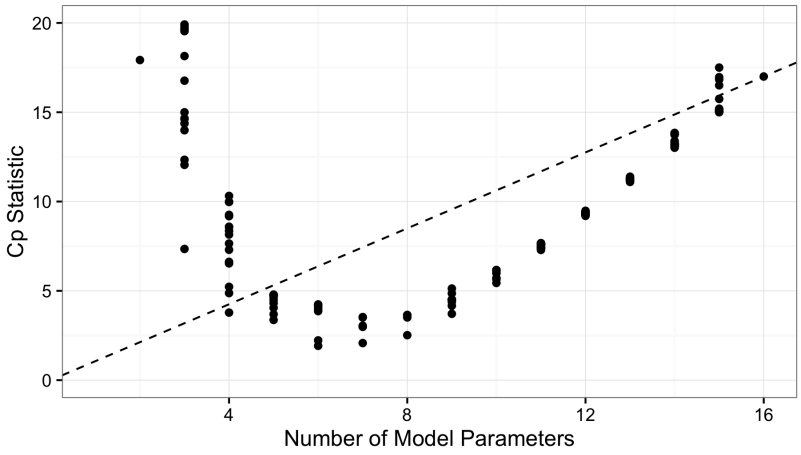

Figure 2: Cp plot for the "good subset models" with relatively low Cp statistics.

Using the results from the 6-variable model and the 9-variable model (including the 6 variables, plus the 3 pollution variables), the data provides convincing evidence that the three pollution variables are associated with mortality (p-value = 0.008313; extra sum of squares F-test).

(b) *Repeat part (a) but use a sequential variable selection technique (forward selection, backward elimination, or stepwise regression). How does the p-value compare?*

Using forward selection, we select a model that includes NonWhite, Educ, JanTemp, House, JulyTemp, Precip, and Density, plus an intercept term.

To this 7-variable model, we add the log-transformed pollution variables and perform an extra-sum-of-squares F-test as shown below.

```
> anova(lmForwardSubsetPoll, lmForwardSubset)
Analysis of Variance Table

Model 1: Mortality ~ NonWhite + Educ + JanTemp + House + JulyTemp + Precip +
    Density + log(HC) + log(NOX) + log(SO2)
Model 2: Mortality ~ NonWhite + Educ + JanTemp + House + JulyTemp + Precip +
    Density
  Res.Df   RSS Df Sum of Sq      F   Pr(>F)
1     49 50403
2     52 63955 -3    -13552 4.3915 0.008162 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Using the results from the 7-variable model and the 10-variable model (including the 7 variables, plus the 3 pollution variables), the data provides convincing evidence that the three pollution variables are associated with mortality (p-value = 0.008162; extra sum of squares F-test). Using forward selection, the p-value is almost identical to (negligibly smaller than) the one found using the model in part (a).

Problem 2: Ramsey 12.20

Problem 3: Ramsey 20.11

Problem 4: Ramsey 20.15

Problem 5:

Problem 6: