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.

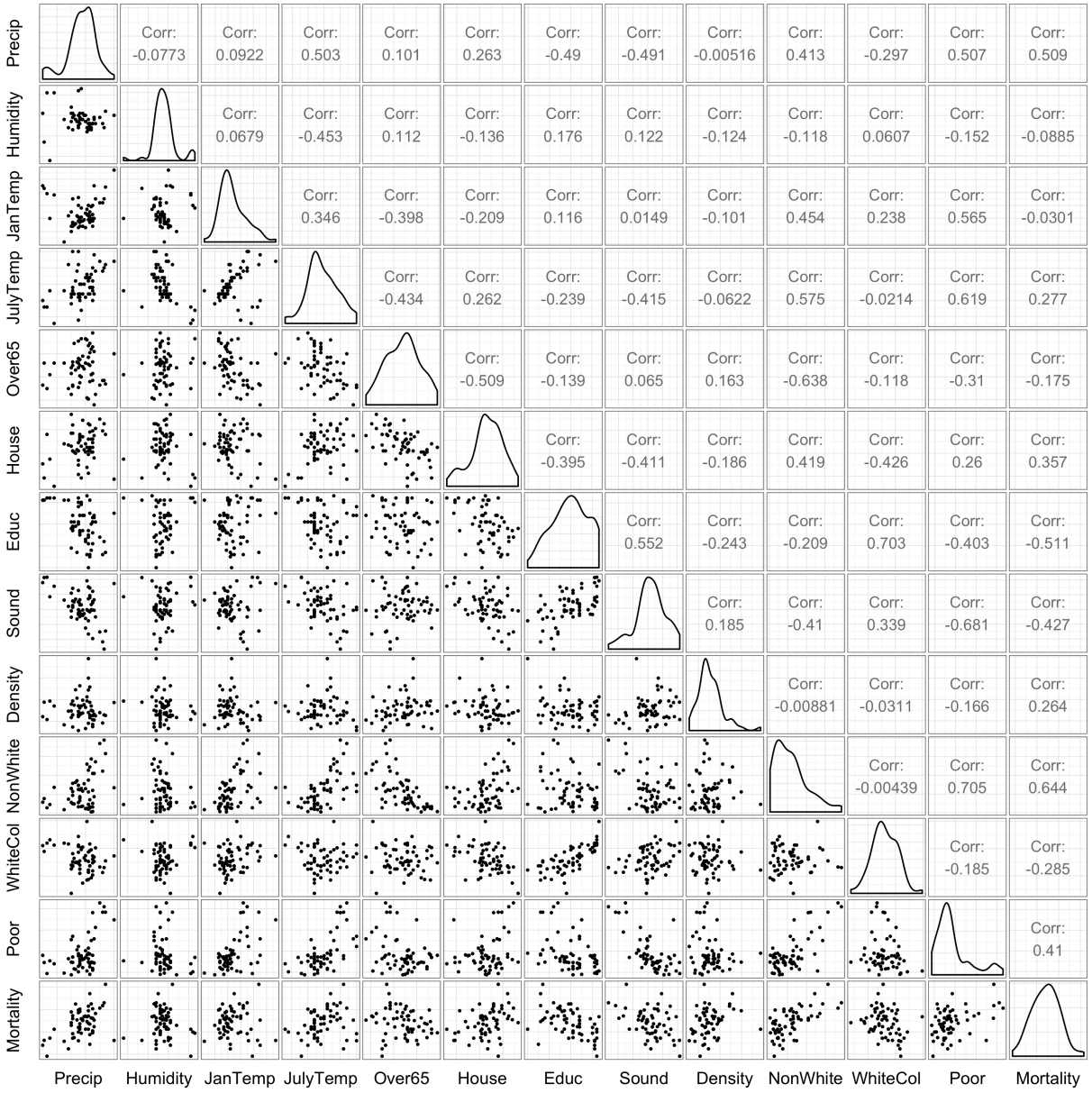

Figure 1: Pairwise scatterplots of the weather and socioeconomic variables from the "Pollution and Mortality" dataset.

```
> 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
```

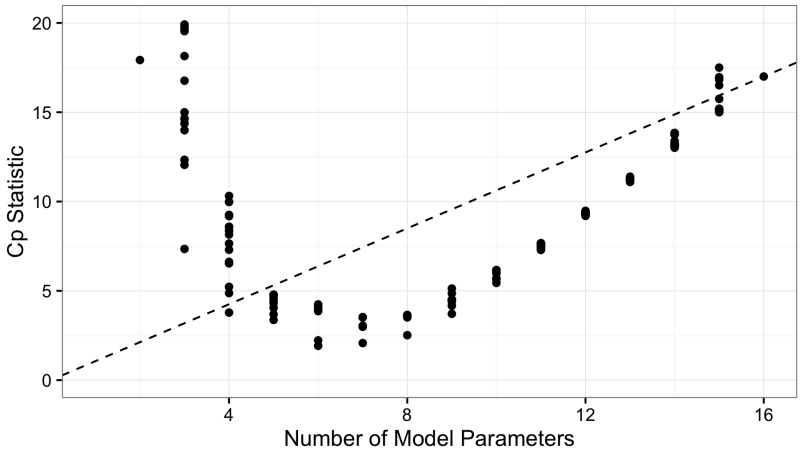

Figure 2: Cp plot for the "good subset models" with relatively low Cp statistics.

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 (just negligibly smaller than) the one found using the model in part (a).

Problem 2: Ramsey 12.20

The number of species on an island is known to be related to the island's area. Of interest is what other variables are also related to the number of species, after island area is accounted for, and whether the answer differs for native species and nonnative species.

We first define the "Nonnative" species value as the total number of species (Total), minus the number of native species (Native). An initial matrix of pairwise scatterplots indicates both (1) severe outliers, and (2) nonlinear behavior between the explanatory variables and Native and Nonnative.

Testing various variable transformations to reduce outliers and make the data more linear, we find that a good set of explanatory variables is log(Area), log(Elev), log(DistNear), log(1 + DistSc), and log(AreaNear). A matrix of pairwise scatterplots with this set of transformed variables is shown in Figure 3. On the untransformed scale, Isabela island has an Area much larger than the rest of the dataset, but on the log scale this is no longer an issue (for neither the Area variable nor the AreaNear variable).

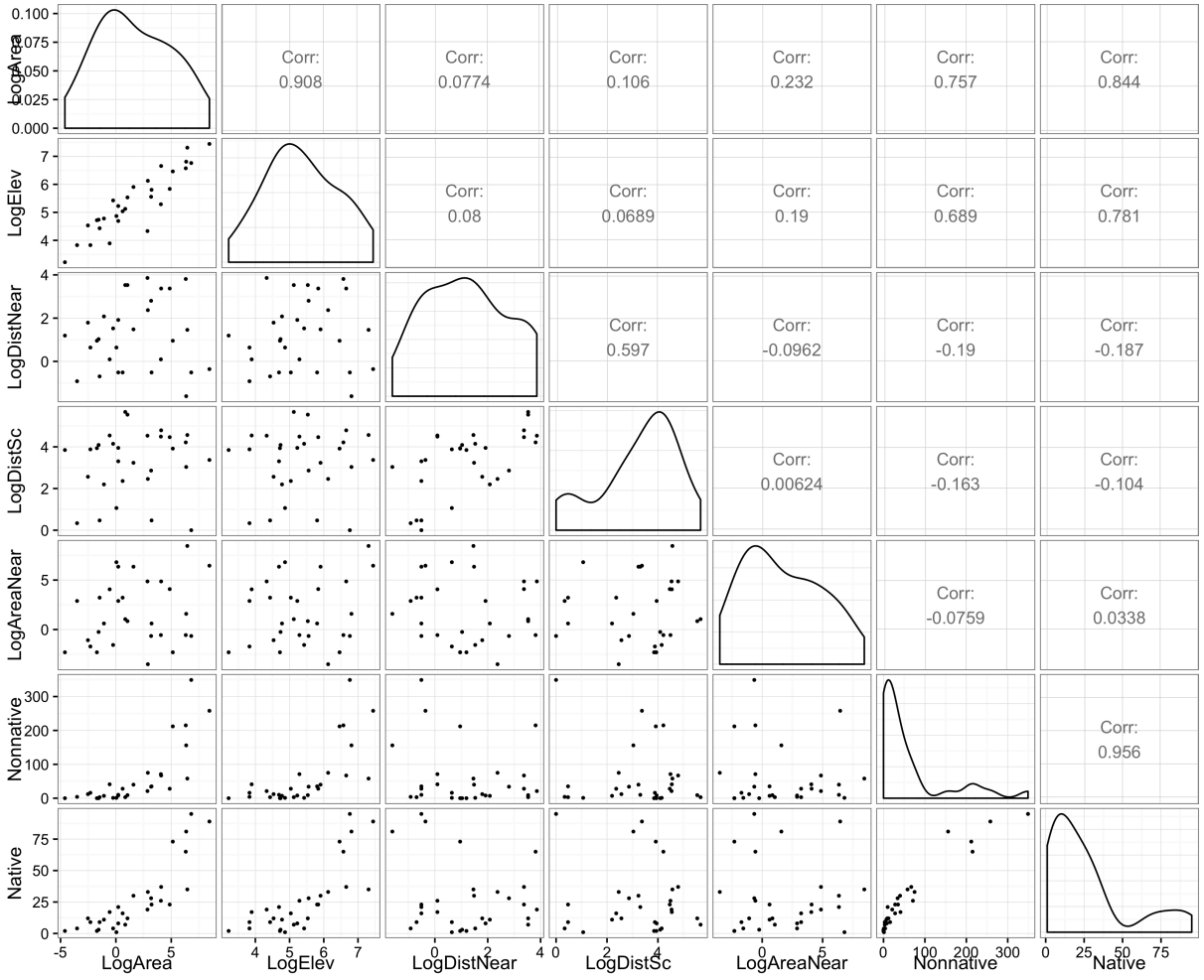

Figure 3: Pairwise scatterplots of the transformed variables in the "Galapagos Islands" dataset.

(a) *Native Species*

A multiple regression of Native on LogArea, LogElev, LogDistNear, LogDistSc, and LogAreaNear is shown below.

```
> # create a linear model with response=Native
> lmNative <- lm(formula = Native ~ LogArea + LogElev +
+                  LogDistNear + LogDistSc + LogAreaNear,
+                data = galapDataTransf)
> summary(lmNative)$coefficients
              Estimate Std. Error    t value     Pr(>|t|)
(Intercept) 17.4186072 26.6342475  0.6539928 0.5193343862
LogArea      6.7877699  1.6800966  4.0401069 0.0004761012
LogElev      1.6306771  5.2281734  0.3119019 0.7578086491
LogDistNear -4.2190775  1.8650490 -2.2621805 0.0330196612
LogDistSc   -0.8665726  1.9225987 -0.4507298 0.6562300815
LogAreaNear -1.6606874  0.7468994 -2.2234419 0.0358553724
```

After accounting for the relationship between Native and Area (where we use the log-transformed value of Area, here), there is suggestive (but inconclusive) evidence that the distance to the nearest island (LogDistNear) is associated with the number of native species on the given island (two sided p-value = 0.0759 for a test that the LogDistNear coefficient is zero). There is moderate evidence that the area of the nearest island (LogAreaNear) is associated with the number of native species on the given island (two sided p-value = 0.0359 for a test that the LogAreaNear coefficient is zero). None of the other variables have a significant relationship with the number of native species.

Examining Cook's Distances for this model in Figure 4, none of the observations are influential.

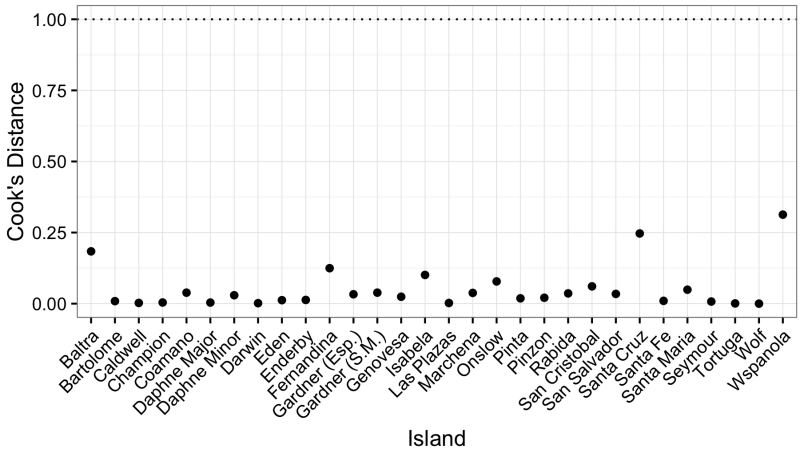

Figure 4: Cook's Distances for each observation in the regression of Native on LogArea, LogElev, LogDistNear, and LogDistSc.

(b) *Nonnative Species*

A multiple regression of Nonnative on LogArea, LogElev, LogDistNear, LogDistSc, and LogAreaNear is shown below.

```
> lmNonnative <- lm(formula = Nonnative ~ LogArea + LogElev +
+                     LogDistNear + LogDistSc + LogAreaNear,
+                   data = galapDataTransf)
> summary(lmNonnative)$coefficients
              Estimate Std. Error    t value  Pr(>|t|)
(Intercept)  90.078403 106.215060  0.8480756 0.4047749
LogArea      23.169532   6.700079  3.4580983 0.0020433
LogElev      -3.250762  20.849500 -0.1559156 0.8774036
LogDistNear -11.513117   7.437653 -1.5479502 0.1347209
LogDistSc    -7.436224   7.667156 -0.9698803 0.3417798
LogAreaNear  -7.974738   2.978570 -2.6773716 0.0131716
```

After accounting for the relationship between Nonnative and Area (where we use the log-transformed value of Area, here), there is convincing evidence that the area of the nearest island (LogAreaNear) is associated with the number of nonnative species on the given island (two sided p-value = 0.0132 for a test that the LogAreaNear coefficient is zero). As compared to the regression of Native, there is no longer any evidence that the distance to the nearest island (LogDistNear) is associated with the number of nonnative species on the given island. None of the other variables have a significant relationship with the number of nonnative species.

Examining Cook's Distances for this model in Figure 5, none of the observations are influential.

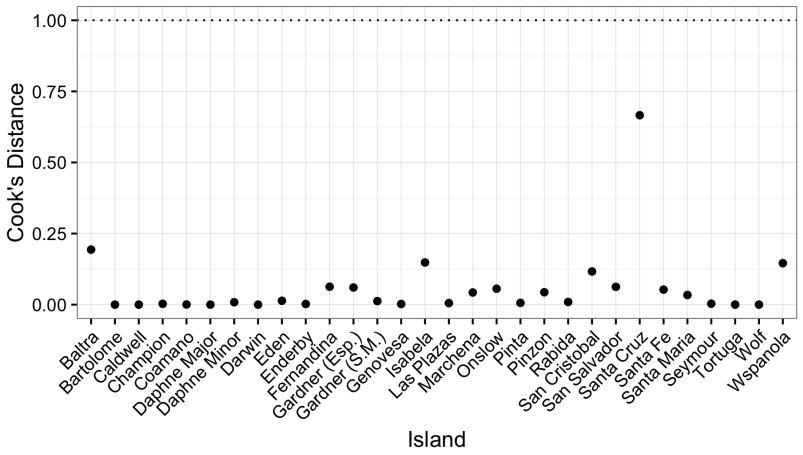

Figure 5: Cook's Distances for each observation in the regression of Nonnative on LogArea, LogElev, LogDistNear, and LogDistSc.

Problem 3: Ramsey 20.11

Problem 4: Ramsey 20.15

Problem 5:

Problem 6: