HUDM6122 Homework_02

Chenguang Pan

2023-02-06

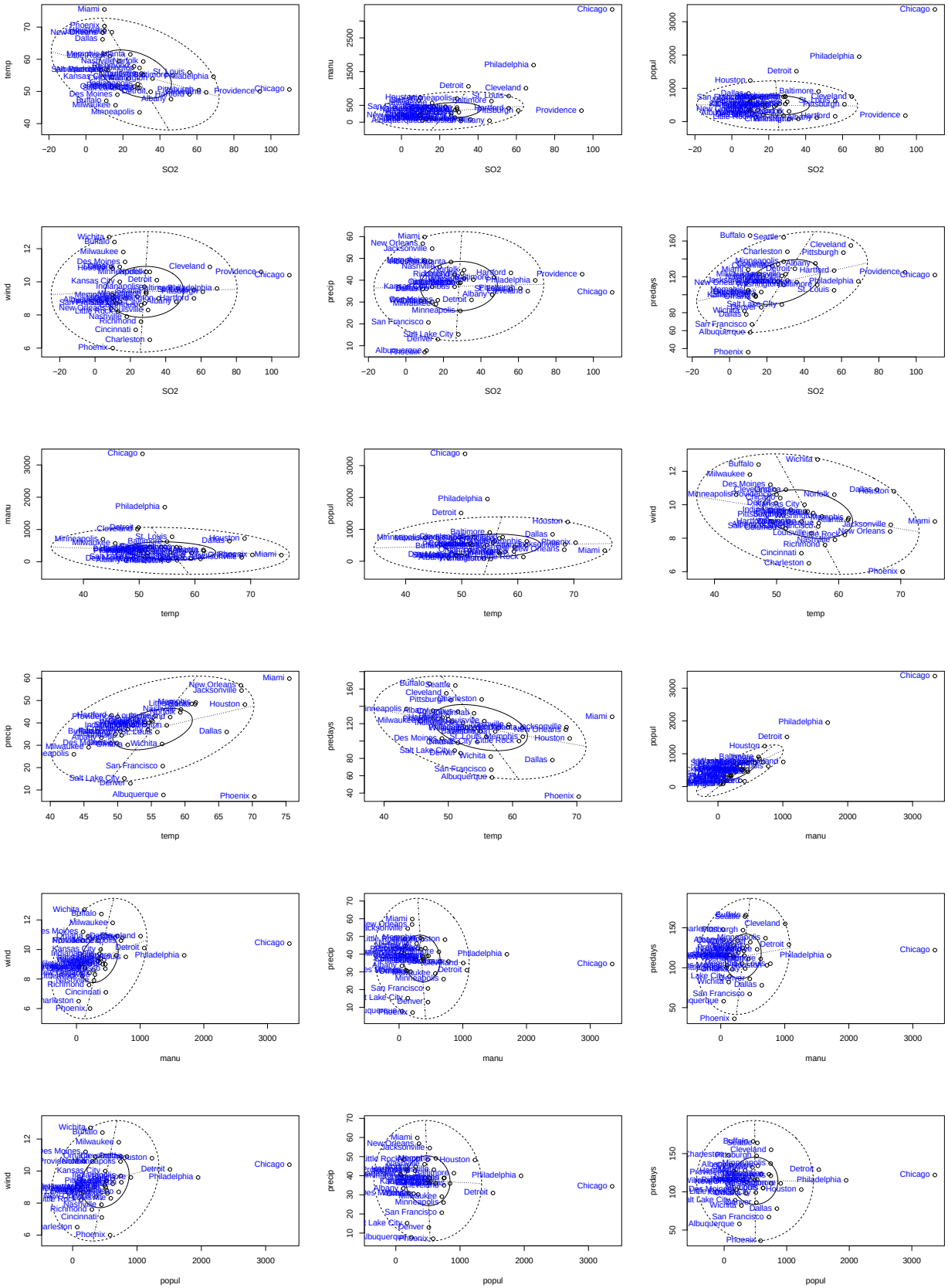
0.1 Ex. 2.1

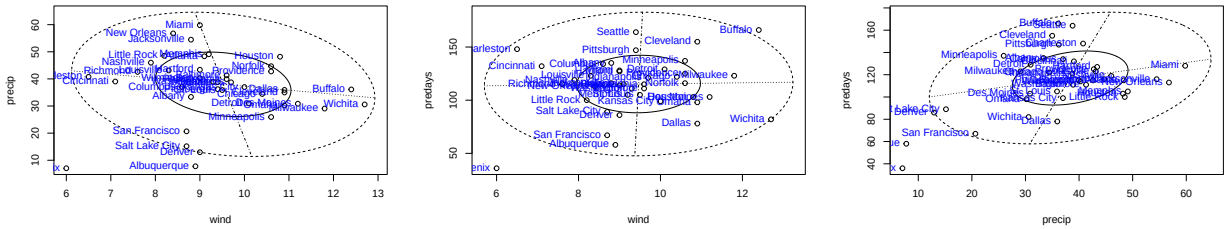
Use the bivariate boxplot on the scatterplot of each pair of variables in the air pollution data to identify any outliers. Calculate the correlation between each pair of variables using all the data and the data with any identified outliers removed. Comment on the results.

MY SOLUTION:

Several techniques worth to be noted: - use a for-loop within a for-loop to map all pairs - use `text()` to give each point a name. - the bivariate boxplot function `bvbox` is inclueded in the package `MVA`

```
> # import the data
> library(HSAUR2)
> library(MVA)
> attach(USairpollution)
> head(USairpollution)
            SO2 temp manu popul wind precip predays
Albany       46 47.6   44   116  8.8  33.36     135
Albuquerque  11 56.8   46   244  8.9   7.77      58
Atlanta      24 61.5  368   497  9.1  48.34     115
Baltimore    47 55.0  625   905  9.6  41.31     111
Buffalo      11 47.1  391   463 12.4  36.11     166
Charleston   31 55.2   35    71  6.5  40.75     148
> dim(USairpollution)
[1] 41  7
> # draw the bivariate boxplot of each pair of variables with a for-loop
> for (i in 1:7) {
+   for (j in i:7) {
+     if (i != j) {
+       var_pair <- USairpollution[, c(i, j)]
+       bvbox(var_pair,
+             xlab = names(USairpollution)[i],
+             ylab = names(USairpollution)[j])
+       text(USairpollution[,i],USairpollution[,j],
+           row.names(USairpollution), # to add the point name
+           pos=2,col = "blue")
+     }
+   }
+ }
```


temp
manu
popul
wind
precip
predays
SO2
temp
manu
popul



From the graphs, we can easily find that the outliers among the observations are “Chicago”, “Detroit”, “Cleveland”,“Philadelphia”, “Miami”, “Phoenix”,“Albuquerque”, “Providence”. I run the correlation matrix on all observations first.

```
> # create the correaltion matrix on all observations
> round(cor(USairpollution),2)
          SO2  temp  manu popul  wind precip predays
SO2      1.00 -0.43  0.64  0.49  0.09   0.05    0.37
temp    -0.43  1.00 -0.19 -0.06 -0.35   0.39   -0.43
manu     0.64 -0.19  1.00  0.96  0.24  -0.03    0.13
popul    0.49 -0.06  0.96  1.00  0.21  -0.03    0.04
wind     0.09 -0.35  0.24  0.21  1.00  -0.01    0.16
precip   0.05  0.39 -0.03 -0.03 -0.01   1.00    0.50
predays  0.37 -0.43  0.13  0.04  0.16   0.50    1.00
> # remove all identified outliers
> drop_city <- match(c("Chicago", "Detroit","Cleveland",
+                      "Philadelphia", "Miami","Phoenix",
+                      "Albuquerque", "Providence"), rownames(USairpollution))
>
> round(cor(USairpollution[-drop_city,]),2)
          SO2  temp  manu popul  wind precip predays
SO2      1.00 -0.40  0.10 -0.16 -0.26  -0.03    0.39
temp    -0.40  1.00 -0.02  0.25 -0.21   0.64   -0.41
manu     0.10 -0.02  1.00  0.82  0.26  -0.15   -0.06
popul   -0.16  0.25  0.82  1.00  0.29   0.03   -0.13
wind    -0.26 -0.21  0.26  0.29  1.00  -0.26   -0.13
precip  -0.03  0.64 -0.15  0.03 -0.26   1.00    0.28
predays  0.39 -0.41 -0.06 -0.13 -0.13   0.28    1.00
```

After dropping all the identified outliers, some of the correlation coefficients has changed to the opposite direction, like from positive to negative, others shrink or increase. It is reasonable since some outliers are with high leverage.

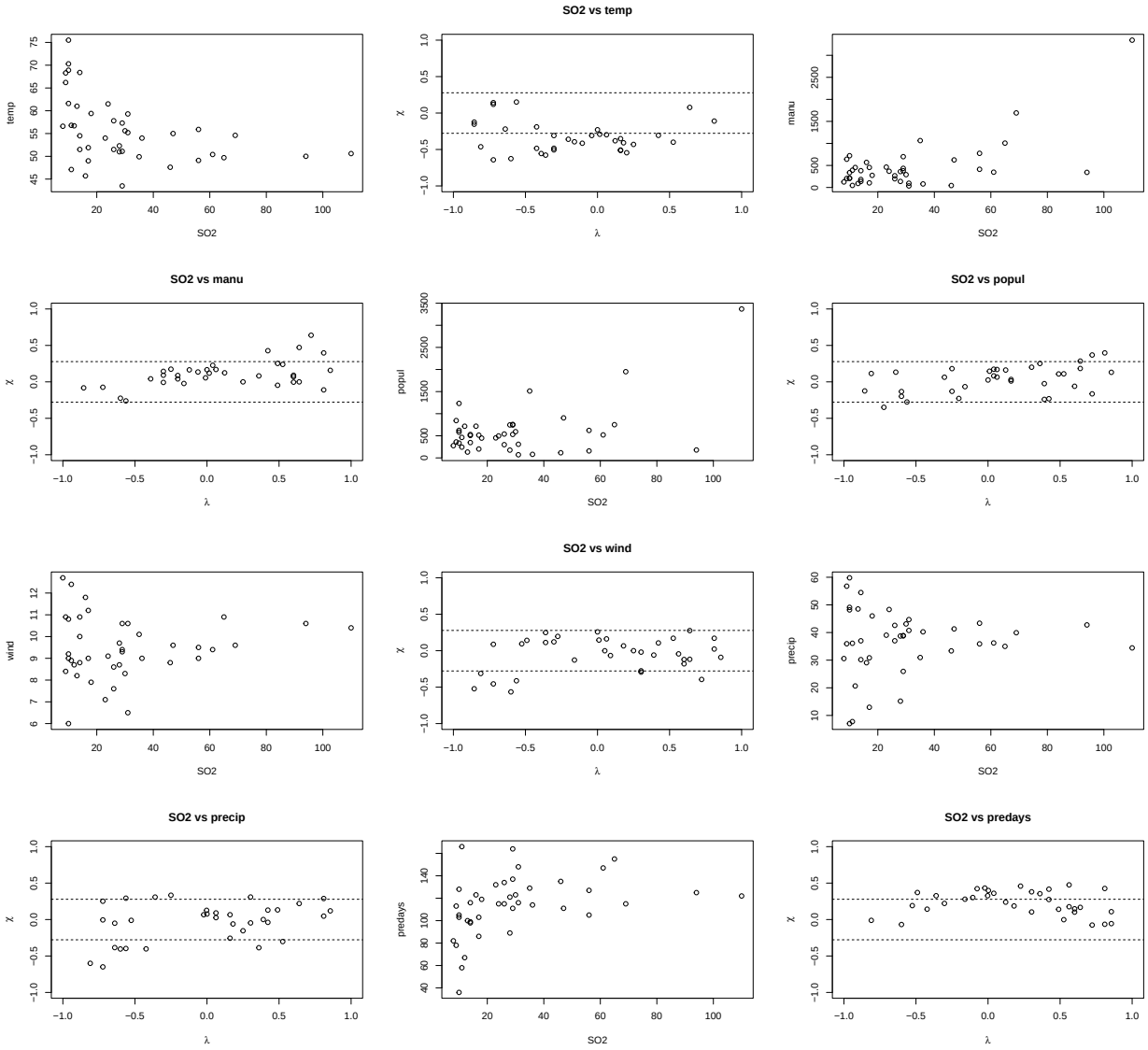
0.2 Ex. 2.2

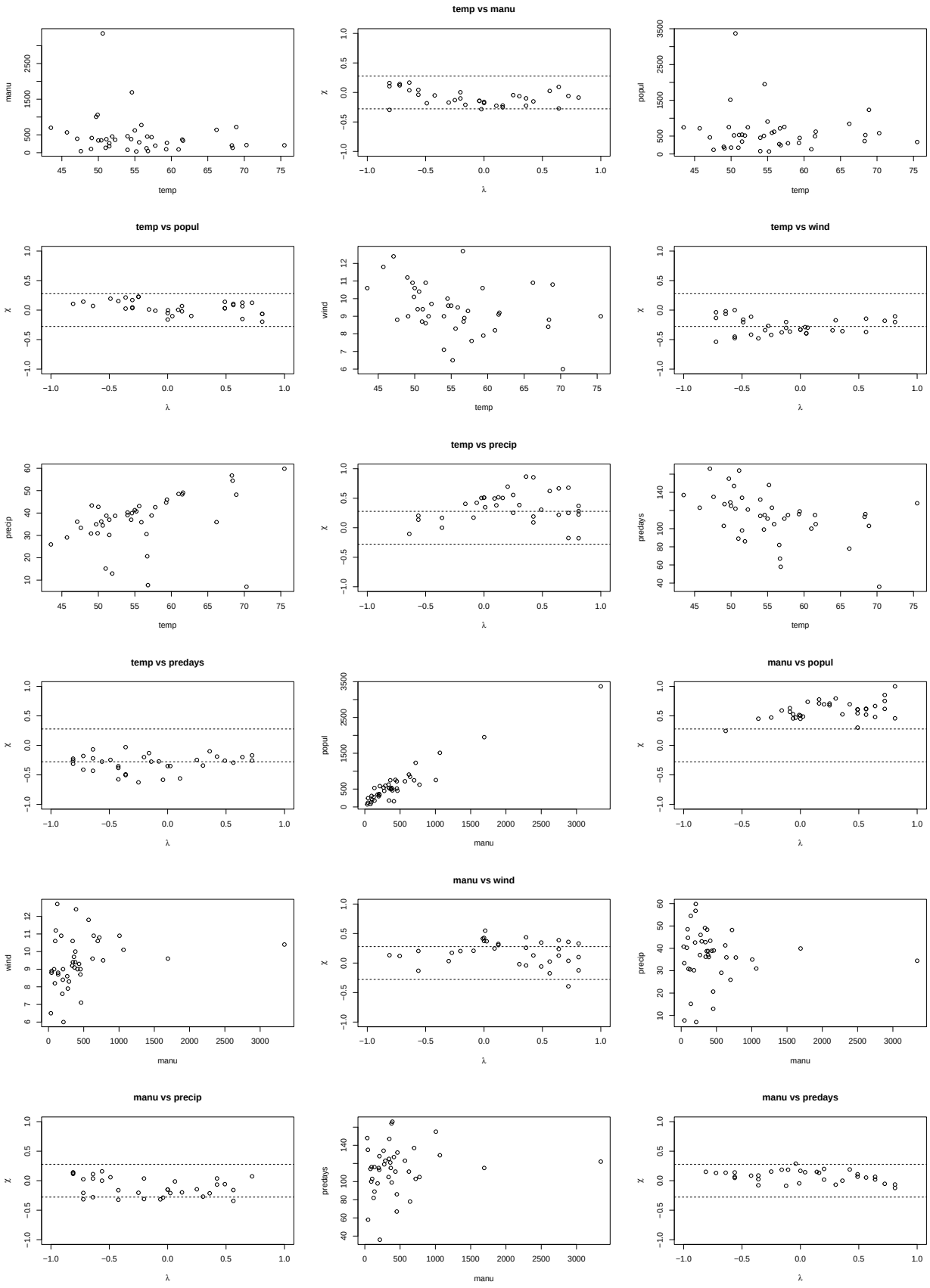
Compare the chi-plots with the corresponding scatterplots for each pair of variables in the air pollution data. Do you think that there is any advantage in the former?

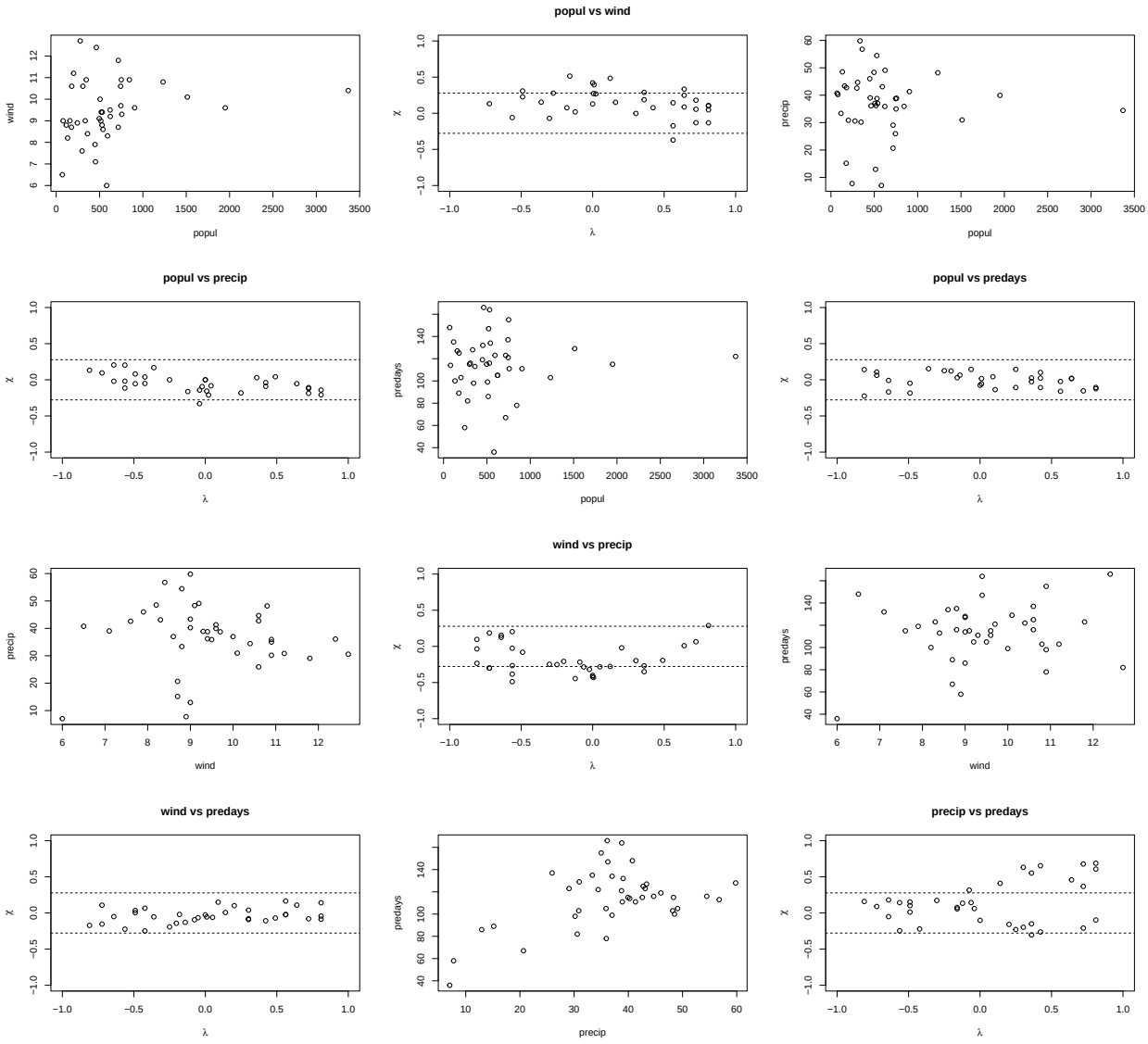
MY SOLUTION:
Several details should be noted.
- For drawing many graphs in `Rmd` file with `knit`, it always reports error or `no such file or directory`. One can clear all the cache in R and cache file and Tex file in the file folder, and do not use the layout function `par()`.
- The `chiplot` function is included in the package `MVA`. If two variables are independent, these value are

asymptotically normal with mean zero; the xi values should show a non-systematic random fluctuation around zero.

```
> for (i in 1:7) {
+   for (j in i:7) {
+     if (i != j) {
+       plot(USairpollution[,i],USairpollution[,j],
+            xlab = names(USairpollution)[i], ylab=names(USairpollution)[j])
+       chiplot(USairpollution[,i],USairpollution[,j],
+               main = paste(names(USairpollution)[i],
+                            "vs",
+                            names(USairpollution)[j]))
+     }
+   }
+ }
```




temp vs manu
manu
temp
χ
λ
popul
temp vs popul
wind
temp vs wind
precip
temp vs precip
predays
temp vs predays
manu vs popul
manu vs wind
manu vs precip
manu vs predays



From the results, one can easily find that the scatter plots are sometimes difficult to identify the independence between two variables. But, comparatively the `chiplot` presents more straightforward way to tell this attribute. For example, it is hard to find the relation from the scattorplot for `manu` and `predays`, but the chiplot clearly demonstrates that these two varriables are independent.

0.3 Ex. 2.3

Compare the chi-plots with the corresponding scatterplots for each pair of variables in the air pollution data. Do you think that there is any advantage in the former?