HUDM6122 Homework_03

Chenguang Pan

2023-02-19

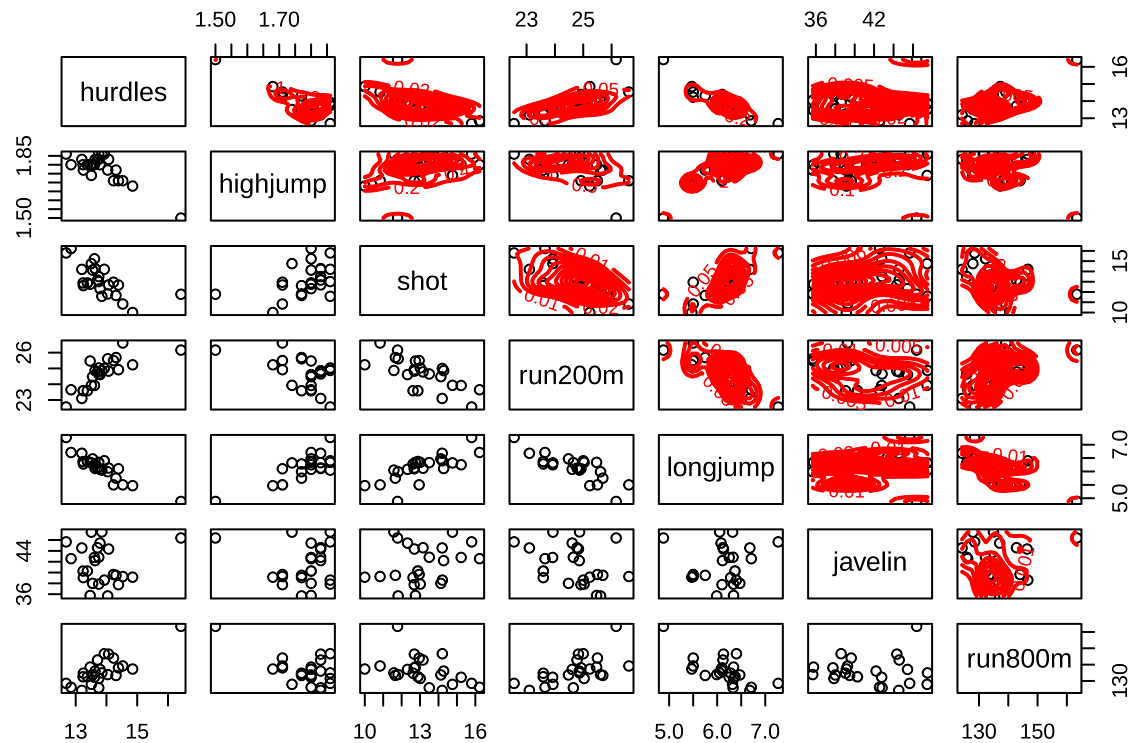
0.1 Ex 3.1

Construct the scatterplot of the heptathlon data showing the contours of the estimated bivariate density function on each panel. Is this graphic more useful than the unenhanced scatterplot matrix?

MY SOLUTION:

Here, I use the `MASS::kde2d()` function to estimate the bivariate density of the data, and plot the contours of the density using the `contour()` function.

```
> # import the package
> library(MVA)
> library(HSAUR2)
> data(heptathlon)
> # Create a scatterplot matrix with density contours
> pairs(heptathlon[, -ncol(heptathlon)], upper.panel = function(x, y) {
+   points(x, y)
+   den <- MASS::kde2d(x, y)
+   contour(den, add = TRUE, col = "red", lwd = 2)})
```



Comparing to the unenhanced scatter plot matrix, this mixed graph can help to easily find the specific characteristics of joint distribution of each pair, like the where is the center of the distribution.

0.2 Ex 3.2

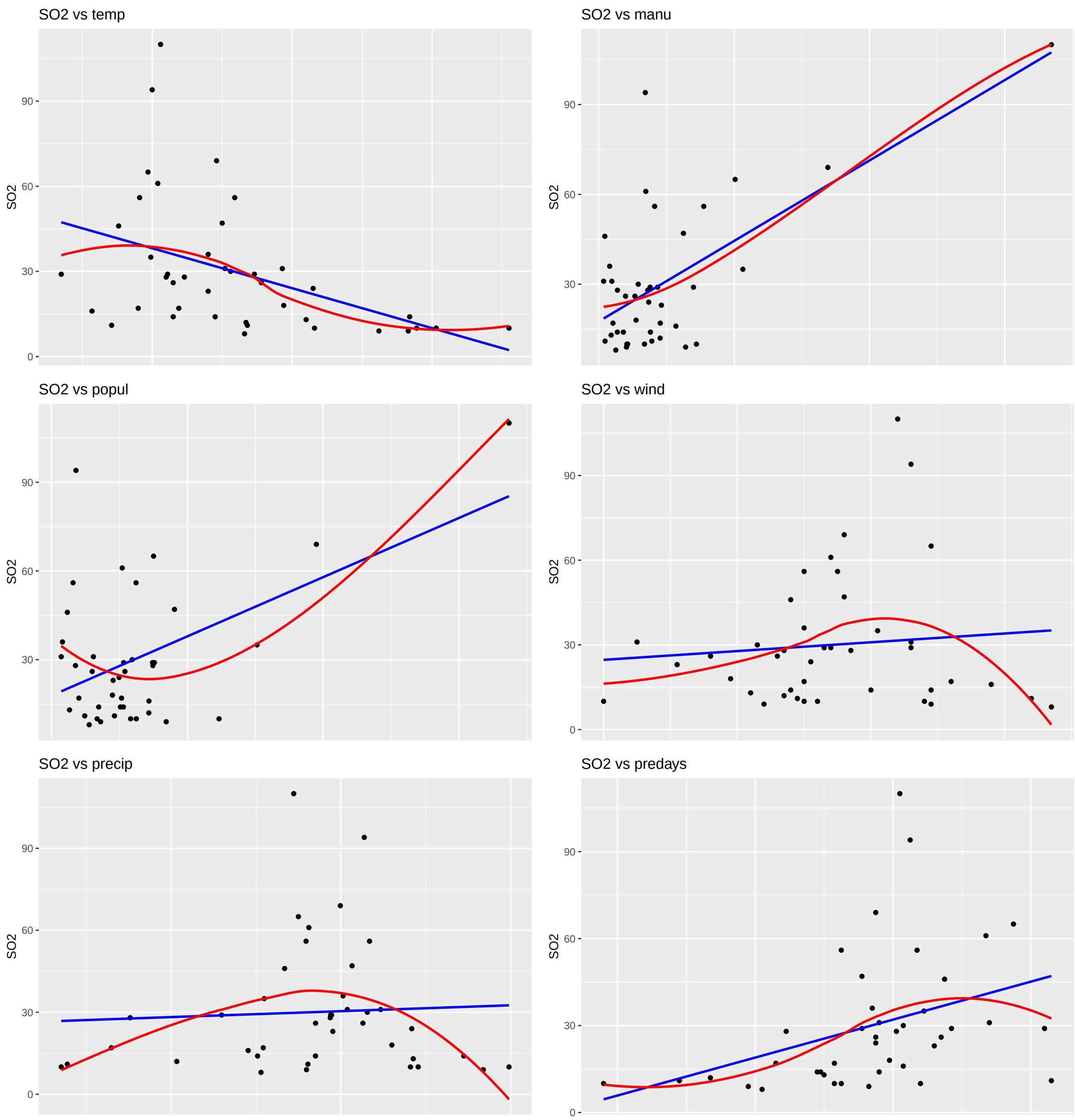
Construct a diagram that shows the SO2 variable in the air pollution data plotted against each of the six explanatory variables, and in each of the scatterplots show the fitted linear regression and a fitted locally weighted regression. Does this diagram help in deciding on the most appropriate model for determining the variables most predictive of sulphur dioxide levels?

MY SOLUTION:

To solve this questions, I used the `ggplot2` to draw each graph.

```
> # try to use a for-loop to get all the maps in fewer lines
> par(mfrow=c(2,3))
> for (i in c(2:7)) {
+   p <- ggplot(USairpollution, aes(x = USairpollution[,i], y = SO2)) +
+                   geom_point() +
+                   geom_smooth(method = "lm", se = FALSE, color = "blue") +
+                   geom_smooth(span = 1, se = FALSE, color = "red") +
+                   labs(title = paste0("SO2 vs ", colnames(USairpollution)[i]))+
+                   # to remove the un-elegant x-axis name
+                   theme(axis.title.x = element_blank(),
+                         axis.text.x = element_blank(),
+                         axis.ticks.x = element_blank())
```

```
+   # assign(var_name, p)
+   print(p)
+ }
```



From six graphs above, we can not easily tell what the strongest predictor is for predicting the SO_2 concentration, since there are some outliers with high leverage in each graph.

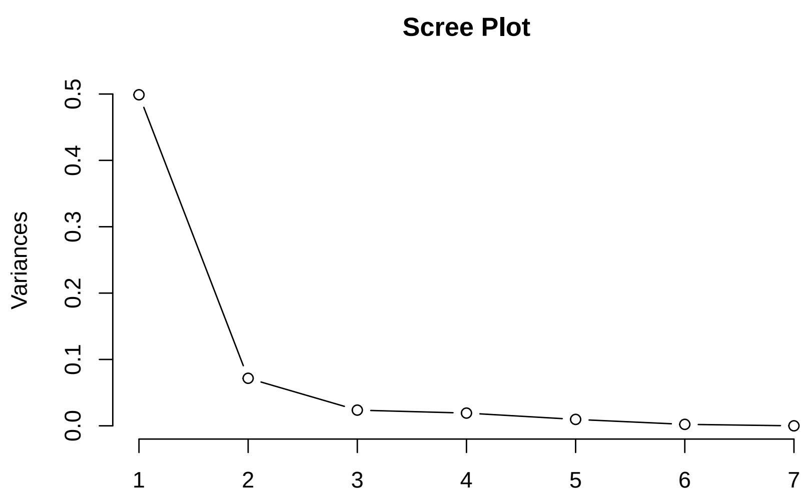
0.3 Ex 3.3

Find the principal components of the following correlation matrix given by MacDonnell (1902) from measurements of seven physical char- acteristics in each of 3000 convicted criminals: How would you interpret the derived components?

MY SOLUTION

First, by using `forceSymmetric()` function in the package `Matrix` to transform the low triangular correlation matrix into a complete correlation matrix. And then, I use `prcomp()` to get the components.

```
> library(Matrix)
> # import the correlation matrix
> corr_lower <- matrix(c(1, 0, 0, 0, 0, 0, 0,
+                        0.402, 1, 0,0,0,0,0,
+                        0.396, 0.618, 1,0,0,0,0,
+                        0.301, 0.150, 0.321, 1,0,0,0,
+                        0.305, 0.135, 0.289, 0.846, 1,0,0,
+                        0.339, 0.206, 0.363, 0.759, 0.797, 1,0,
+                        0.340, 0.183, 0.345, 0.661, 0.800, 0.736, 1),7,7, byrow = T)
> # generate a complete correlation matrix
> corr_symmetric <- forceSymmetric(corr_lower, uplo="L")
> # use prcomp to calculate the principal components
> pca <- prcomp(corr_symmetric, scale. = FALSE)
> # get the PCA results
> summary(pca)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6       PC7
Standard deviation     0.7063 0.2677 0.15376 0.13851 0.09832 0.04686 3.458e-17
Proportion of Variance 0.7979 0.1146 0.03781 0.03069 0.01546 0.00351 0.000e+00
Cumulative Proportion  0.7979 0.9125 0.95034 0.98103 0.99649 1.00000 1.000e+00
> # draw the scree plot
> plot(pca, type = "l",
+      main = "Scree Plot")
```



The first two components can explain 91.25% variance of the total. Therefore, I will choose to use the first two components to represent this dataset.