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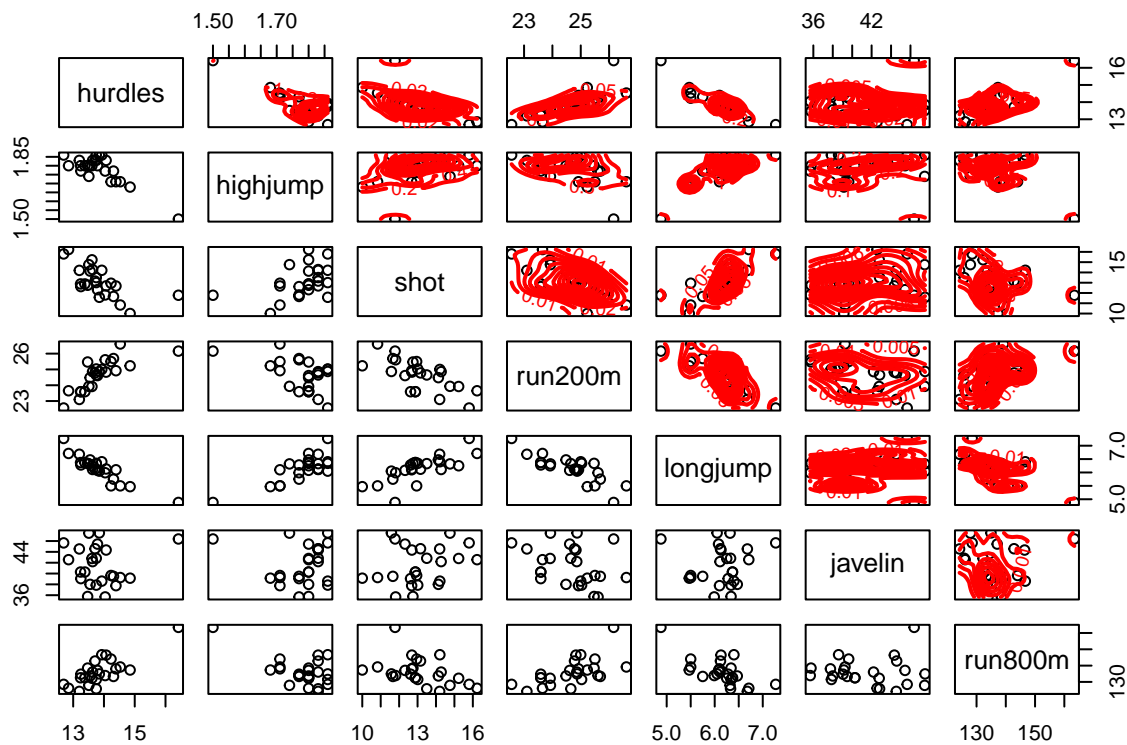
Comparing to the unenhanced scatter plot matrix, this mixed graph can help to easily find the specific characteristics of joint distribution of each pair, like the where is the center of the distribution.

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)
+ }
```
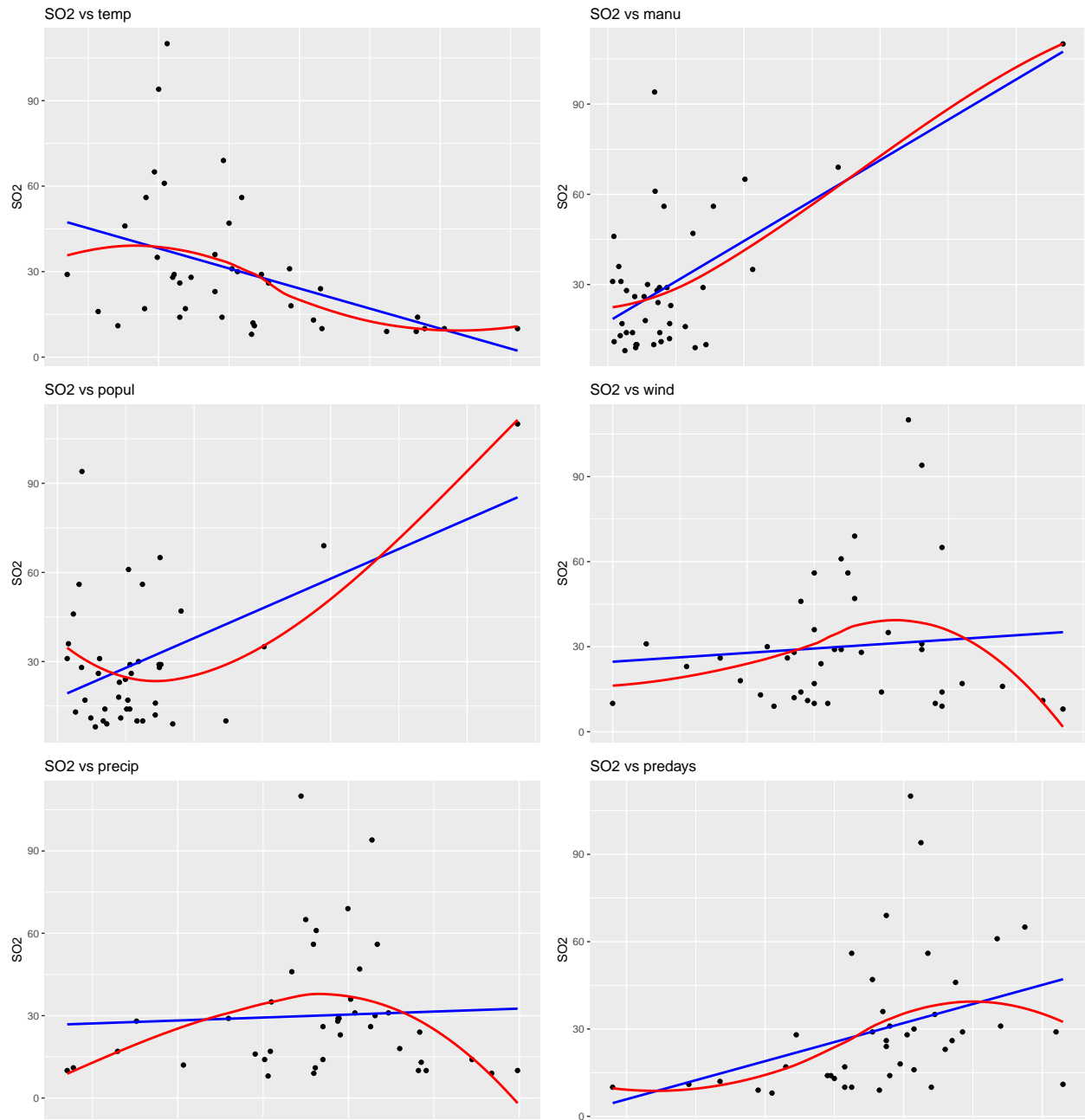

From six graphs above, we can not easily tell what the strongest predictor is for predicting the SO_2 concentration, since there are some outliers with high leverage in each graph.