HW3

Getong Zhong

2023-02-24

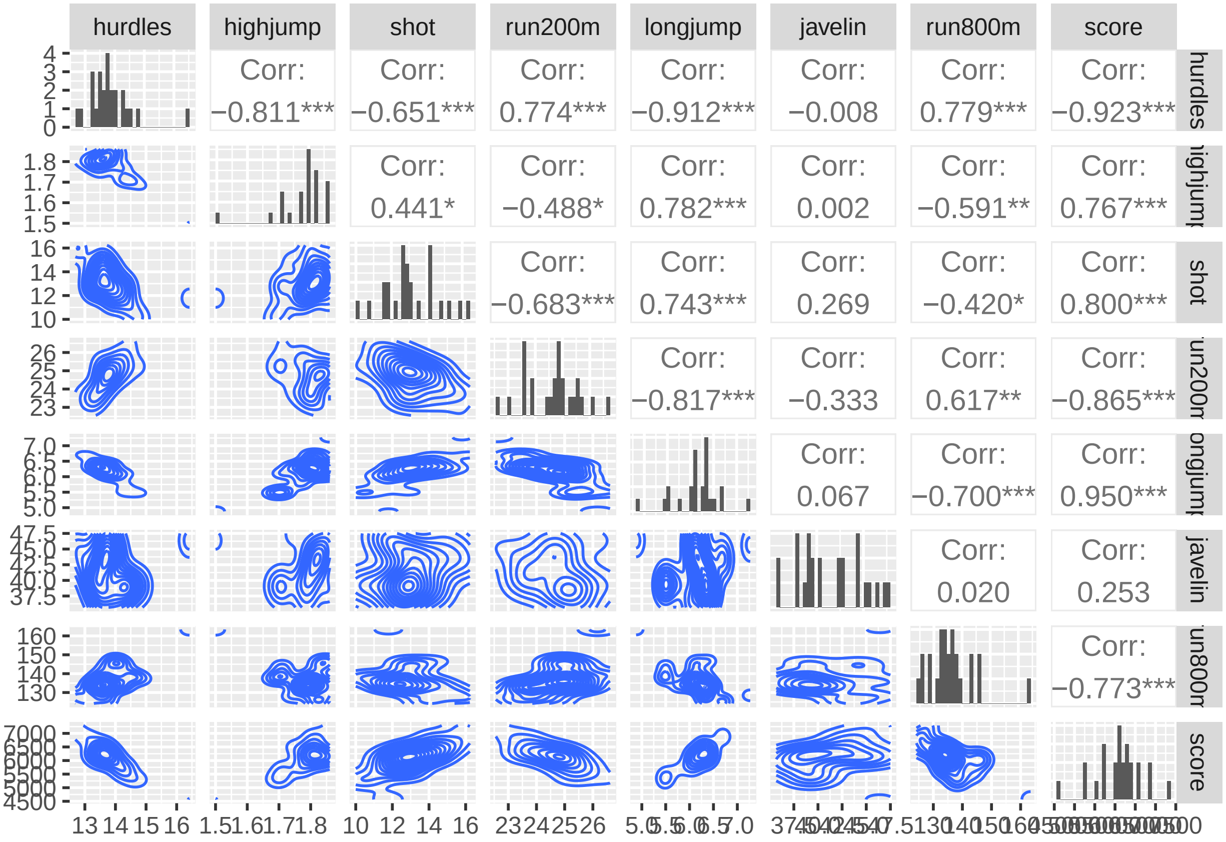
3.1

From the below graph, a enhanced version of scatter plot matrix showing the estimated bivarite densities on each panel, we can know not only the relationship between the two variables but also have an eye on the estimated bivariate density on each panel of the data.

```
data("heptathlon", package = "HSAUR2")
ggpairs(heptathlon,
        lower = list(continuous = "density"),
        diag = list(continuous = "barDiag", binwidth = 30),
        upper = list(continuous = "cor"), size = 5)
```

```
## Warning in warn_if_args_exist(list(...)): Extra arguments: "size" are being
## ignored. If these are meant to be aesthetics, submit them using the 'mapping'
## variable within ggpairs with ggplot2::aes or ggplot2::aes_string.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



3.2

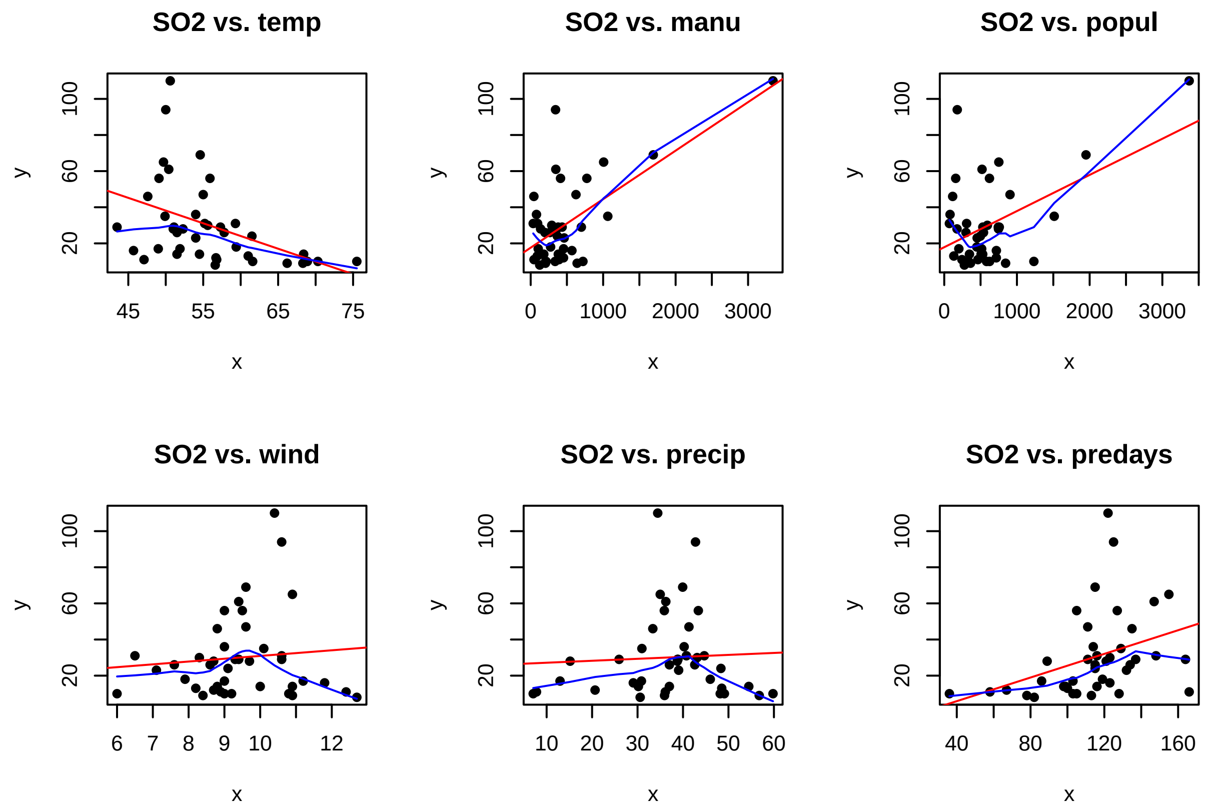
From the graph below I'm not sure which variable is the most predictive of SO2 since some of the have the similar graph results, as they all not have an explict linear relation with the variable SO2, plotting them out can only find out their relationship between the SO2 but not deciding if they are predictive in the model. Compare the others, temp and predays has the most power prediction model with SO2.

```
data("USairpollution", package = "HSAUR2")

scatterplot <- function(x, y, ...) {
  plot(x, y, pch = 16, col = "black", ...)
  abline(lm(y ~ x), col = "red")
  lines(stats::lowess(x, y), col = "blue")
}


par(mfrow = c(2, 3))

scatterplot(USairpollution$temp, USairpollution$SO2, main = "SO2 vs. temp")
scatterplot(USairpollution$manu, USairpollution$SO2, main = "SO2 vs. manu")
scatterplot(USairpollution$popul, USairpollution$SO2, main = "SO2 vs. popul")
scatterplot(USairpollution$wind, USairpollution$SO2, main = "SO2 vs. wind")
scatterplot(USairpollution$precip, USairpollution$SO2, main = "SO2 vs. precip")
scatterplot(USairpollution$predays, USairpollution$SO2, main = "SO2 vs. predays")
```



## 3.3 From the principal components we get, in the first PC, we can see that the highest loading are 0.45 and 0.5 which are associate with variable “Left finger length” and “Left forearm length” and the lowest loading is -0.44 from variable “Head breadth”, that is from the first component of PC, we know that the first principal component can be interpreted as a linear combination of the original variables that captures the variation in the “Left finger length” and “Left forearm length” while minimizing the influence of the “Head breadth”. Other components can be exaplained in the similar way.

```
cor <- matrix(c(1, 0.402, 0.396, 0.301,0.305, 0.339, 0.34,0.402,1,0.618, 0.15, 0.135, 0.206, 0.183, 0.39

pca <- prcomp(cor)

pca
```

```
## Standard deviations (1, .., p=7):
## [1] 6.781627e-01 2.670446e-01 2.122651e-01 1.524801e-01 1.384780e-01
## [6] 4.687054e-02 5.267718e-17
##
## Rotation (n x k) = (7 x 7):
##             PC1         PC2         PC3         PC4         PC5          PC6
## [1,] -0.1680649  0.80652466 -0.19813343  0.23020282 -0.05753937  0.073405874
## [2,] -0.4412697 -0.31764201 -0.06543239 -0.53414988  0.12783539  0.060971809
## [3,] -0.2724569 -0.46795327 -0.13505927  0.77869004 -0.12768383  0.100325036
## [4,]  0.4498397 -0.09953724  0.15377505 -0.04296196 -0.61440145 -0.461122364
## [5,]  0.4969729 -0.06742908  0.15803625 -0.02257381 -0.03371484  0.815017403
## [6,]  0.3014564 -0.11310157 -0.93886876 -0.11507088 -0.01125957 -0.003459013
## [7,]  0.4032228 -0.04893829  0.09013310  0.19933834  0.76504350 -0.322381354
##              PC7
## [1,]  0.46938553
## [2,]  0.62827353
## [3,]  0.23626147
## [4,]  0.41487485
## [5,]  0.23999192
## [6,] -0.03843782
## [7,]  0.31294630
```

3.4

```
data("frets", package = "boot")
headsize <- frets

depr <- c(
 0.212,
 0.124,  0.098,
-0.164,  0.308,  0.044,
-0.101, -0.207, -0.106, -0.208,
-0.158, -0.183, -0.180, -0.192, 0.492)
LAdepr <- diag(6) / 2
LAdepr[upper.tri(LAdepr)] <- depr
LAdepr <- LAdepr + t(LAdepr)
rownames(LAdepr) <- colnames(LAdepr) <- c("CESD", "Health", "Gender", "Age", "Edu", "Income")
x <- LAdepr
LAdepr <- as.data.frame(LAdepr)

headsize_cancor <- cancor(headsize[,1:2], headsize[,3:4])

headsize_eigenvals <- eigen(cov(headsize))$values

LAdepr_cancor <- cancor(LAdepr[,1:3], LAdepr[,4:6])

LAdepr_eigenvals <- eigen(cov(LAdepr))$values

headsize_test_stat <- (-1) * (nrow(headsize) - 1 - 0.5 * (2 + 2 + 1)) * sum(log(1 - headsize_cancor$cor

headsize_pval <- pchisq(headsize_test_stat, df = 2*2, lower.tail = FALSE)

LAdepr_test_stat <- (-1) * (nrow(LAdepr) - 1 - 0.5 * (3 + 3 + 1)) * sum(log(1 - LAdepr_cancor$cor^2))


LAdepr_pval <- pchisq(LAdepr_test_stat, df = 3*3, lower.tail = FALSE)

cat("Headsize data: test statistic =", headsize_test_stat, "p-value =", headsize_pval)
```

```
## Headsize data: test statistic = 20.96418 p-value = 0.0003218897
```

```
cat("Depression data: test statistic =", LAdepr_test_stat, "p-value =", LAdepr_pval)
```

```
## Depression data: test statistic = 54.96492 p-value = 1.236845e-08
```

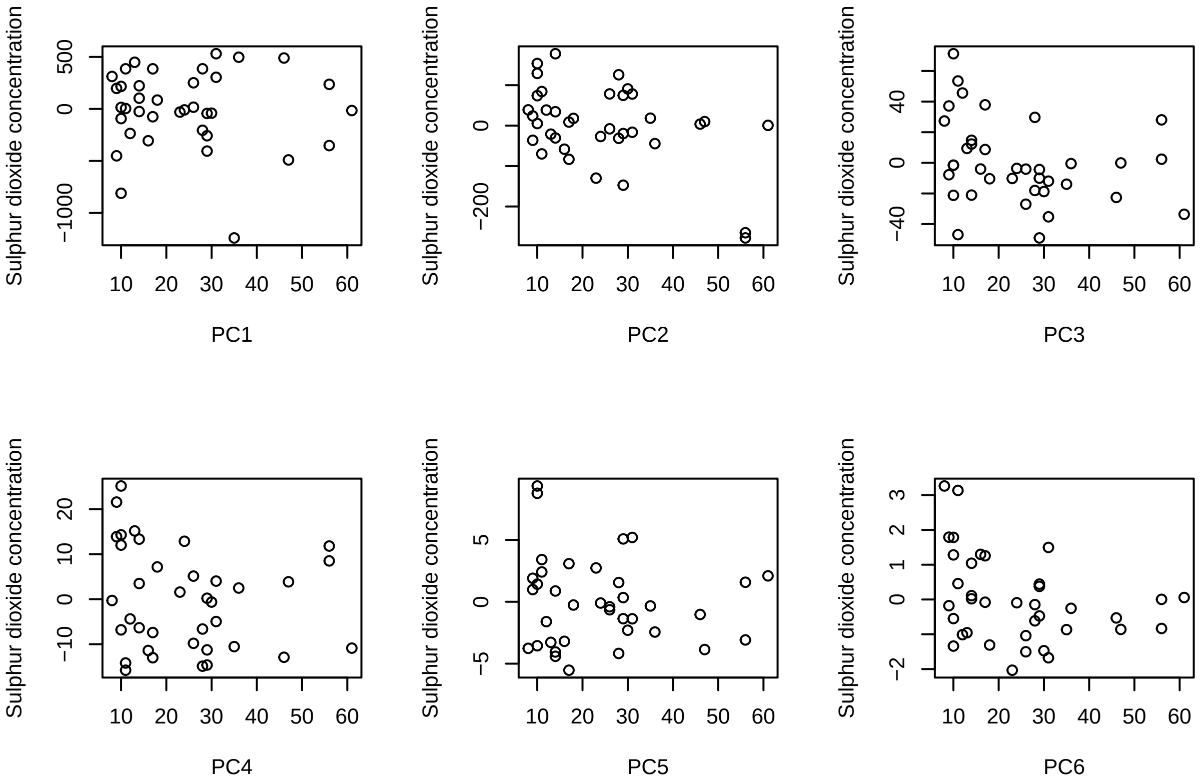
3.5

Except PC1, other PCs vs SO2 all has a weak negative relationship. Except PC1 from the plot we can see that the points are not loosely distributed which indicate a strong co-relationship, although might not be linear, SO2 is strongly related to the variation captured by that component. This suggests that the SO2 is an important driver of that component and should be considered in any analysis or modeling that involves that component.

```r
data("USairpollution", package = "HSAUR2")
USairpollution <- USairpollution[-c(7, 9, 30, 33), ]
airpollution_pca <- princomp(USairpollution[,-1])

airpollution_scores <- airpollution_pca$score

par(mfrow = c(2, 3))
out <- sapply(1:6, function(i) {
     plot(USairpollution$SO2,airpollution_scores[,i],
          xlab = paste("PC", i, sep = ""),
          ylab = "Sulphur dioxide concentration")
     })
```


Sulphur dioxide concentration
PC1
Sulphur dioxide concentration
PC2
Sulphur dioxide concentration
PC3
Sulphur dioxide concentration
PC4
Sulphur dioxide concentration
PC5
Sulphur dioxide concentration
PC6