

Midterm Test for ST4240 Data Mining

(please answer all the questions for full marks. Please send your answer to staxyc@nus.edu.sg)

1. For [data A](http://www.stat.nus.edu.sg/~staxyc/DM07testdata1.dat) (at <http://www.stat.nus.edu.sg/~staxyc/DM07testdata1.dat>), there are 5 predictors X_1, \dots, X_5 and response Y . A Single-index model (SIM) is suggested

$$Y = g(a_1 X_1 + \dots + a_5 X_5) + e$$

- A. Estimate the model, plot the link function and its confidence band.

The estimated model is

$$Y = g(0.008595073X_1 - 0.740091476X_2 + 0.034182754X_3 - 0.671579756X_4 + 0.001703434X_5)$$

The estimated function and its 95% confidence band are show in Figure 1

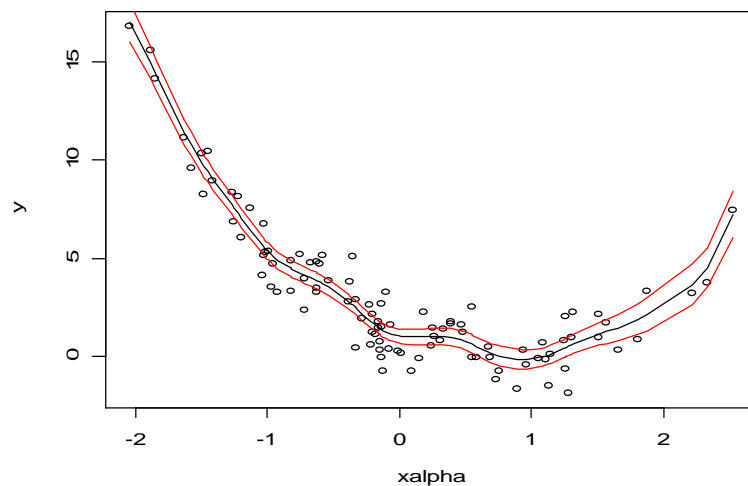


Figure 1

- B. which variables can be removed? Estimate the model again after removing the variables

The estimated coefficients have SE respectively 0.02222250 0.01297481 0.02339194 0.01451642 0.02307927. By checking the “t-statistics”, we can see that X_1, X_3 and X_5 can be removed

- C. For a new X ($X_1=0, X_2=0, X_3 = 0, X_4=0, X_5=0$), predict the function value i.e. $E(Y|\text{new } X)$ and calculate its 95% confidence interval.

Predict value is 1.023846, the 95% confidence interval is [0.6610302, 1.386661]

CODE

```
xy = read.table("testdata1.dat")
x = data.matrix(xy[,1:5])
y = data.matrix(xy[,6])

source("sim.R")
out = sim(x, y)

out$alpha
out$se

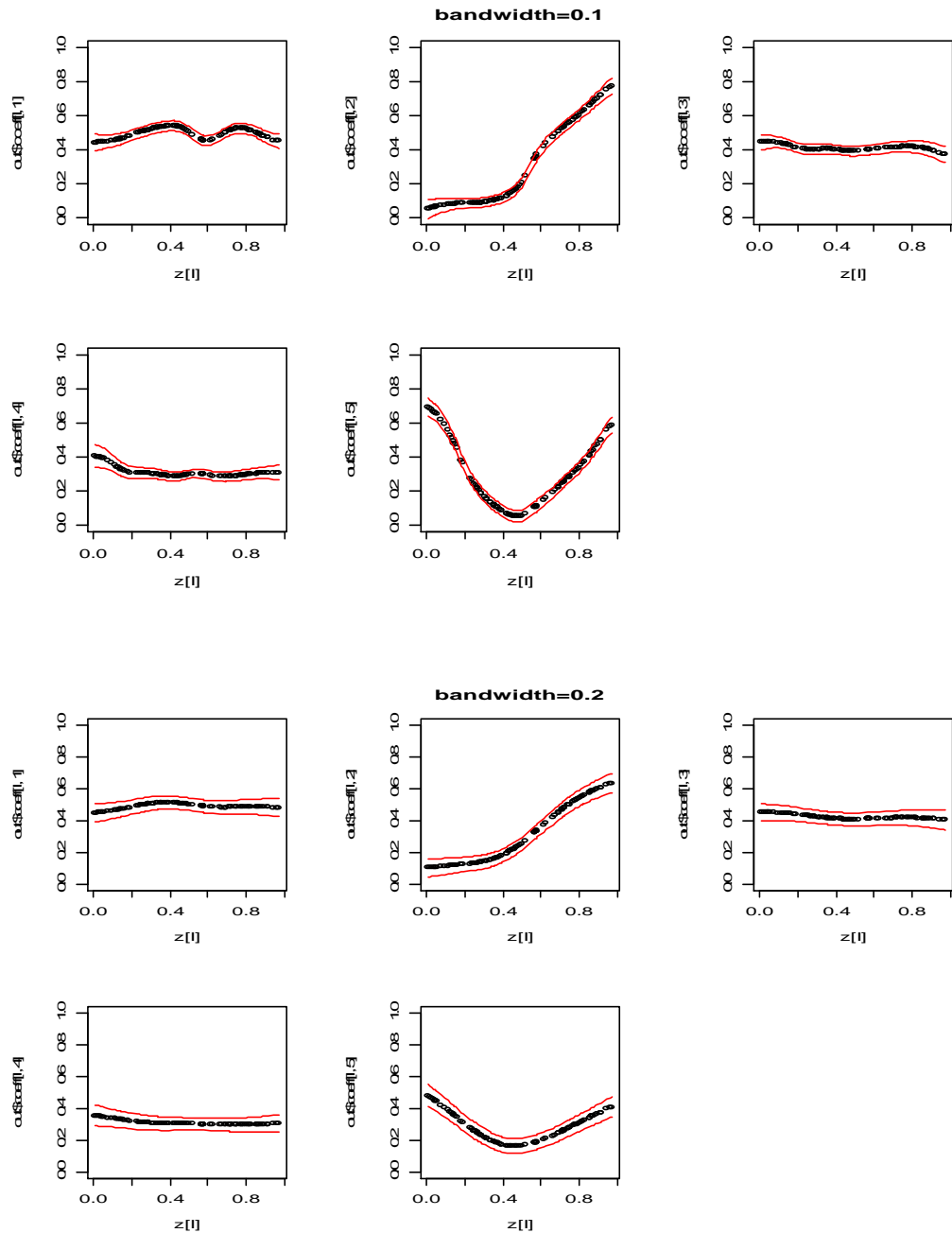
xalpha = x %*% out$alpha
plot(xalpha, y)
I = order(xalpha)
lines(xalpha[I], out$predict[I])
lines(xalpha[I], out$Ln[I])
lines(xalpha[I], out$Un[I])

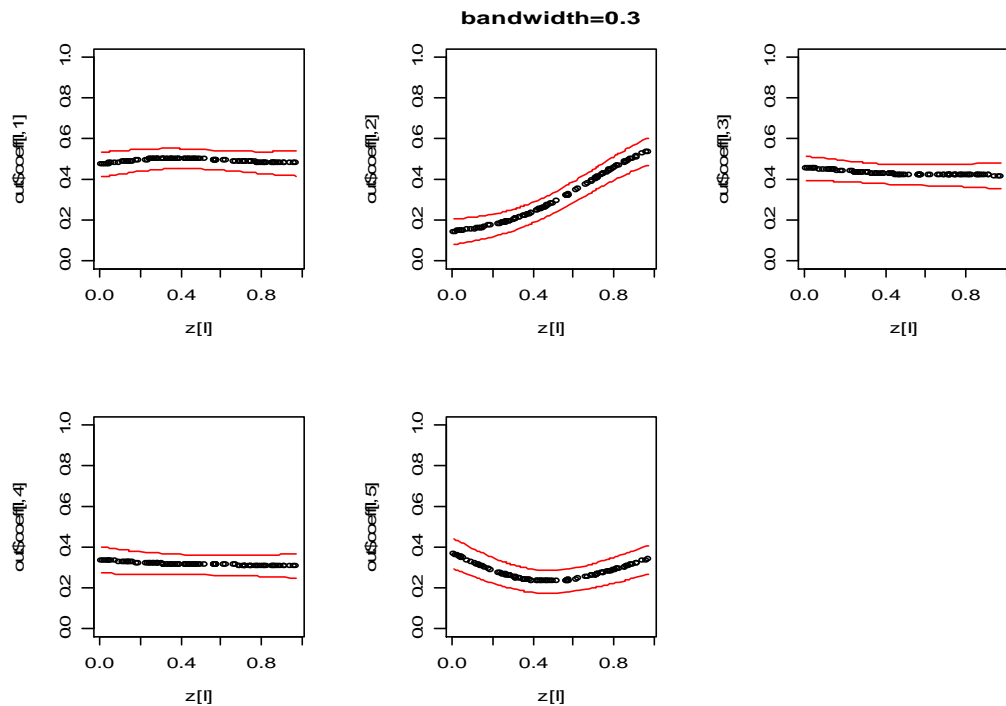
out = sim(x, y, xnew = c(0, 0, 0, 0, 0))
out$predict
out$Ln
out$Un
```

2. For [data B](http://www.stat.nus.edu.sg/~staxyc/DM07testdata2.dat) (at <http://www.stat.nus.edu.sg/~staxyc/DM07testdata2.dat>), there are 5 predictors X1, ..., X4 and Z with response Y. To find how Z affects the relationships between Y and X1, ..., X4 we consider the varying coefficient model

$$Y = g_0(Z) + g_1(Z)*X_1 + g_2(Z)*X_2 + g_3(Z)*X_3 + g_4(Z)*X_4 + e$$

- A. try different bandwidths 0.1, 0.2, 0.3, which bandwidth do you prefer?
 We try the model with the 3 bandwidths. The plots are shown below





By comparing the plots, it suggests that $h=0.1$ is overfitted (undersmoothed) and $h=0.3$ is underfitted (oversmoothed); $h=0.2$ is suggested

- B. which coefficient functions among g_0, g_1, \dots, g_4 are constant and which are varying? [hint: using `plot(..., ylim=c(specify, specify))`
 $a_0(\cdot), a_2(\cdot)$ and $a_3(\cdot)$ are constants
- C. for a new observation with $X_1 = 1, X_2 = 1, X_3=0.5, X_4 = 0.5$ and $Z = 0$, predict its Y
the predicted Y is 1.426397

CODE

```
xy = read.table("testdata2.dat")
x = data.matrix(xy[,1:4])
z = data.matrix(xy[,5])
y = data.matrix(xy[,6])

source("vcm.R")
#####
out = vcm(x, z, y, bandwidth=0.2)
I = order(z)
par(mfrow = c(2, 3))
plot(z[I], out$coeff[I,1], ylim=c(0, 1))
lines(z[I], out$Ln[I,1], col="red")
lines(z[I], out$Un[I,1], col="red")

plot(z[I], out$coeff[I,2], ylim=c(0, 1))
```

```
lines(z[I], out$Ln[I,2], col="red")
lines(z[I], out$Un[I,2], col="red")
title("bandwidth=0.2")

plot(z[I], out$coeff[I,3], ylim=c(0, 1))
lines(z[I], out$Ln[I,3], col="red")
lines(z[I], out$Un[I,3], col="red")

plot(z[I], out$coeff[I,4], ylim=c(0, 1))
lines(z[I], out$Ln[I,4], col="red")
lines(z[I], out$Un[I,4], col="red")

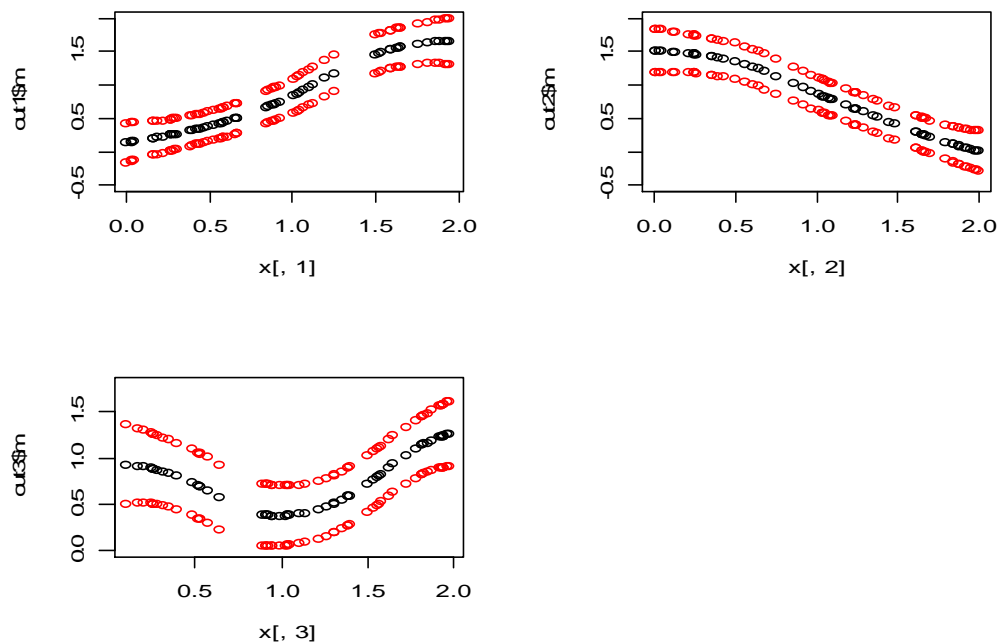
plot(z[I], out$coeff[I,5], ylim=c(0, 1))
lines(z[I], out$Ln[I,5], col="red")
lines(z[I], out$Un[I,5], col="red")

out = vcm(x, z, y, znew = 0, bandwidth=0.2)
predict = out$coeff[1] + out$coeff[2] + out$coeff[3] +0.5*out$coeff[4]+
0.5*out$coeff[5]
predict
```

3. For [data C](http://www.stat.nus.edu.sg/~staxyc/DM07testdata3.dat) (at <http://www.stat.nus.edu.sg/~staxyc/DM07testdata3.dat>), there are 3 predictors X1, X2, X3 and response Y.

A. plot Y against each covariate X1, X2 and X3 respectively, and the regression function $m_1(x) = E(Y|X_1=x)$, $m_2(x) = E(Y|X_2=x)$, $m_3(x) = E(Y|X_3=x)$ and corresponding 95% confidence bands

The estimated functions are shown below



B. Suppose we need to construct a partially linear model, what is your suggestion? i.e, which predictor you select as the nonlinear part?

From the above plot, we see that the relation between Y and X1 is likely linear, Y and X2 is likely linear, with X3 nonlinear. We would choose X3 as the nonlinear part in the partially linear model

C. For X1 = 0, X2 = 2, X3 = 0.5, predict its response Y.

The predicted value is -1.090471

CODE

```
xy = read.table("testdata3.dat")
x = data.matrix(xy[,1:3])
y = data.matrix(xy[,4])

source("cvh.R")
source("ks.R")

par(mfrow = c(2, 2))
h = cvh(x[,1], y)
out1 = ks(x[,1], y, bandwidth = h)
plot(x[,1], out1$m, ylim=c(-0.5,2))
```

```
lines(x[,1], out1$L, type="p", col="red")
lines(x[,1], out1$U, type="p", col="red")

h = cvh(x[,2], y)
out2 = ks(x[,2], y, bandwidth = h)
plot(x[,2], out2$m, ylim=c(-0.5,2))
lines(x[,2], out2$L, type="p", col="red")
lines(x[,2], out2$U, type="p", col="red")

h = cvh(x[,3], y)
out3 = ks(x[,3], y, bandwidth = h)
plot(x[,3], out3$m, ylim=c(0,1.8))
lines(x[,3], out3$L, type="p", col="red")
lines(x[,3], out3$U, type="p", col="red")

source("plr.R")
out = plr(x[,1:2], x[,3], y, znew=0.5, bandwidth=h);
predict = out$beta[1]*0 + out$beta[2]*2 + out$g
predict
```