Measurement Theory
Spring 2015
Lecturer: Christopher Fariss

Problem Set 1

01/29/2015

```
library(knitr, quietly = TRUE)
opts_chunk$set(echo=FALSE)
opts_chunk$set(tidy=FALSE)
library(edarf, quietly = TRUE)
library(ggplot2, quietly = TRUE)
library(dplyr, quietly = TRUE)
library(randomForest, quietly = TRUE)
```

Figure 1: Bivariate Scatterplots with Y

]

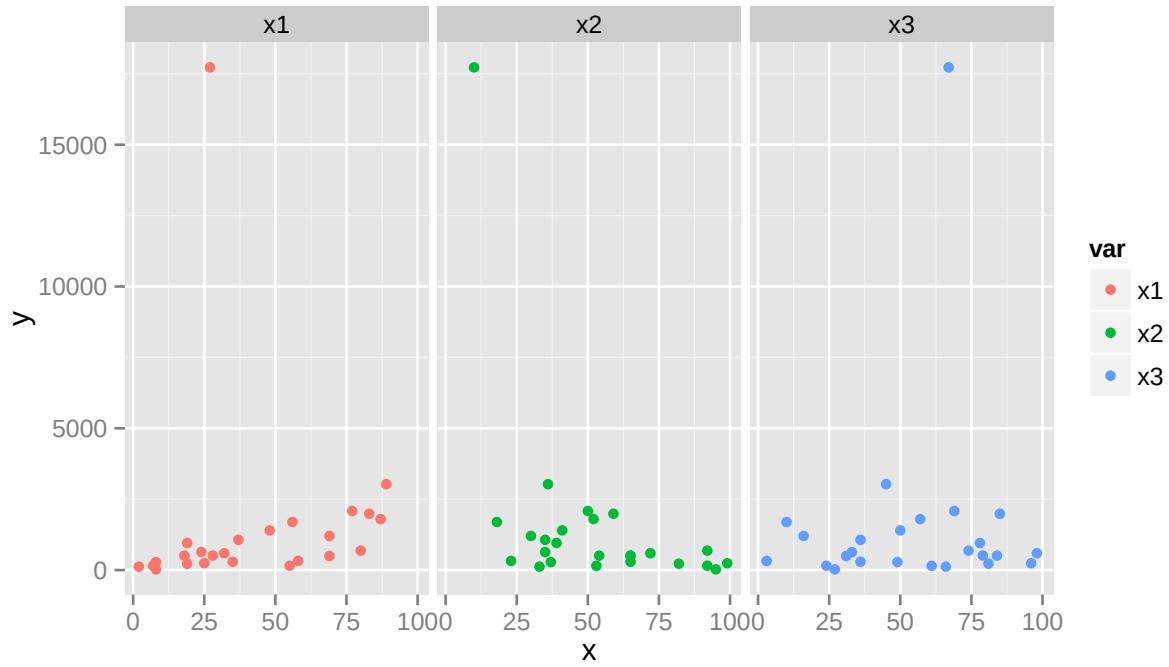
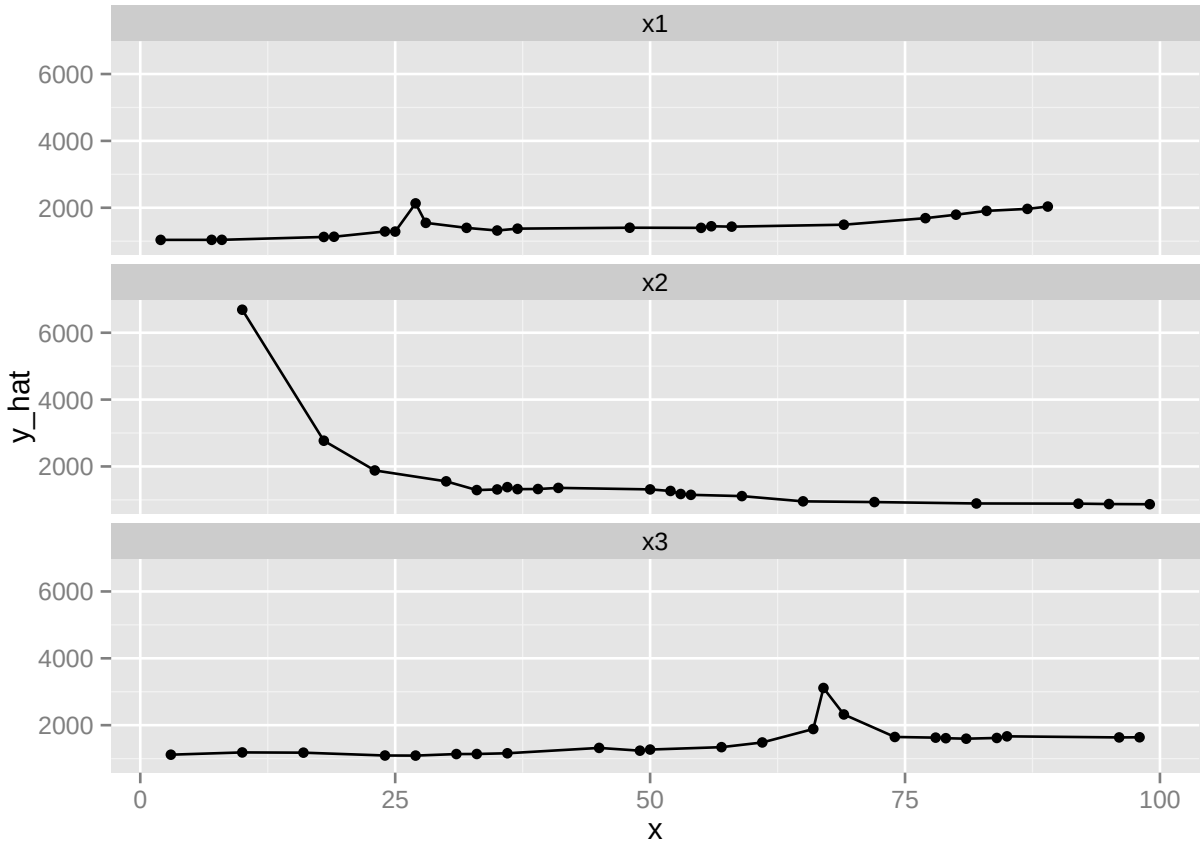


Figure 1 displays the bivariate scatterplots. There might be a relationship between x_1 and x_2 and y respectively. But everything is definitely masked by the one outlier in y.

Figure 2 shows partial dependence plots for $x1$, $x2$ and $x3$. The only intersting thing are the spikes that are due to the outlier in y.

Figure 2: Partial Dependence Plot from Random Forest

]



1 R Code

```
library(knitr, quietly = TRUE)
opts_chunk$set(echo=FALSE)
opts_chunk$set(tidy=FALSE)
library(edarf, quietly = TRUE)
library(ggplot2, quietly = TRUE)
library(dplyr, quietly = TRUE)
library(randomForest, quietly = TRUE)
setwd("~/Dropbox/Spring2015/PLSC597B_Measurement/hw1")
dat <- read.table('_data_problem_Set_01.csv', sep = ",", header = T)

# Scatterplot with each x
pdat <- data_frame(x = c(dat$x1, dat$x2, dat$x3), y = rep(dat$y, 3),
                   var = rep(c("x1", "x2", "x3"), each = nrow(dat)))
p <- ggplot(data = pdat, aes(x, y))
p <- p + geom_point(aes(x, y, color = var))
p <- p + facet_wrap( ~ var)
```

```
p
# Look at correlations
fit <- randomForest(y ~ x1 + x2 + x3, data = dat, importance = T)
imp <- as.data.frame(fit$importance)
imp$var <- rownames(imp)
pd1 <- partial_dependence(fit, dat, c("x1"), cutoff = 22)
pd2 <- partial_dependence(fit, dat, c("x2"), cutoff = 21)
pd3 <- partial_dependence(fit, dat, c("x3"), cutoff = 24)
colnames(pd1) <- colnames(pd2) <- colnames(pd3) <- c("x", "y_hat")
pd <- rbind(pd1, pd2, pd3)
pd$var <- c(rep("x1", 22), rep("x2", 21), rep("x3", 24))

p <- ggplot(pd, aes(x, y_hat))
p <- p + facet_wrap(~ var, ncol = 1)
p <- p + geom_line()
p <- p + geom_point()
p
```