Homework 1 - Zachary Jones

```
df <- read.csv("hw_01.csv")

library(ggplot2)
library(reshape2)

ggplot(melt(df, id.vars = c("unit", "y")), aes(value, y, group = variable)) +
    geom_point() + facet_wrap(~variable, scales = "free") + geom_smooth(method = "loess")
```

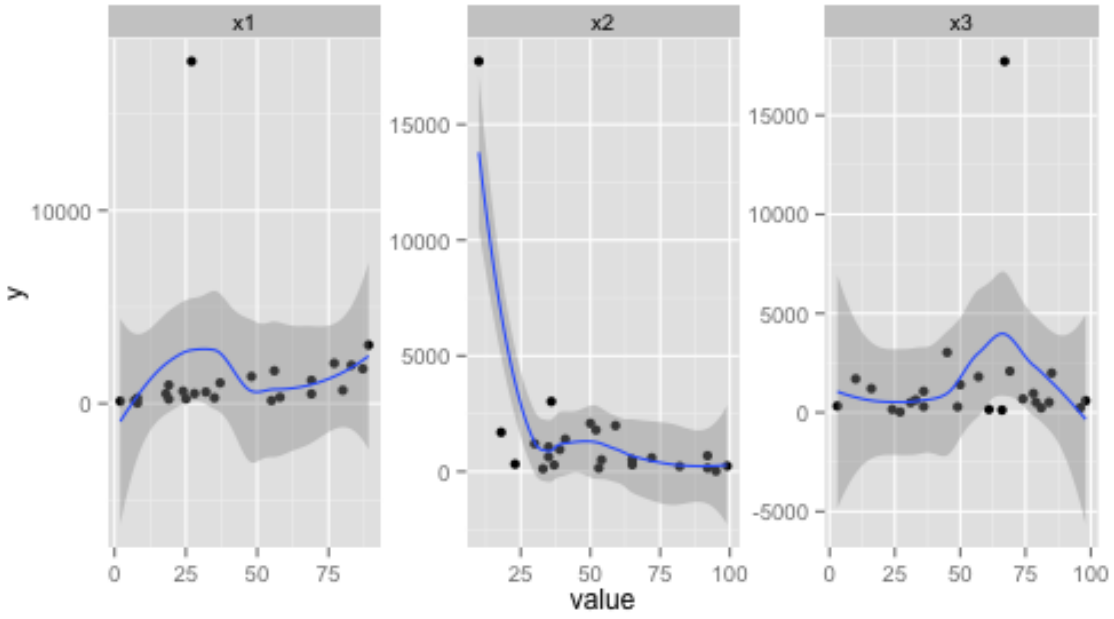

Figure 1: plot of chunk bivariate loess

So far it doesn't look like there is much of a relationship between any of the $\mathbf{x}_j$ and $\mathbf{y}$. There is one outlier on all three of the bivariate plots which is way higher on $\mathbf{y}$ than any of the other points in its neighborhood on each of the $\mathbf{x}_j$.

```
ggplot(melt(df, id.vars = "unit"), aes(value)) + geom_histogram() + facet_wrap(~variable,
    scales = "free")
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

Looking at the empirical distribution of the variables, the $\mathbf{x}_j$ look nearly uniformly distributed on $[0, 100]$ and $\mathbf{y}$ looks as if it were draws from a fat-tailed non-negative distribution. With such little data though this might be quite wrong.

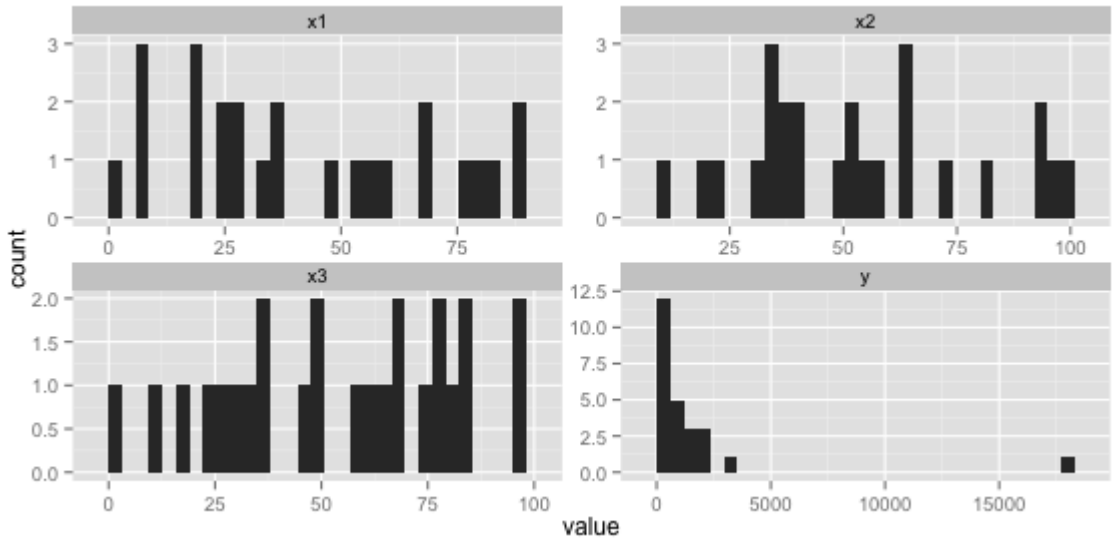

Figure 2: plot of chunk histograms

```
library(party)

## Loading required package: grid
## Loading required package: mvtnorm
## Loading required package: modeltools
## Loading required package: stats4
## Loading required package: strucchange
## Loading required package: zoo
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
##
## Loading required package: sandwich

fit <- cforest(y ~ x1 + x2 + x3 + unit, df, controls = cforest_control(mtry = 2,
    ntree = 1000))
mean((predict(fit) - df$y)^2)

## [1] 10862840

mean(abs(predict(fit) - df$y))

## [1] 1408.431

out <- data.frame(label = c("x1", "x2", "x3", "unit"), `marginal importance` = varimp(fit),
    `conditional importance` = varimp(fit, conditional = TRUE), check.names = FALSE)
```

```
ggplot(melt(out, id.vars = "label"), aes(label, value)) + geom_point() + geom_hline(xintercept = 0,
    linetype = "dotted") + facet_grid(~variable, scales = "free")
```

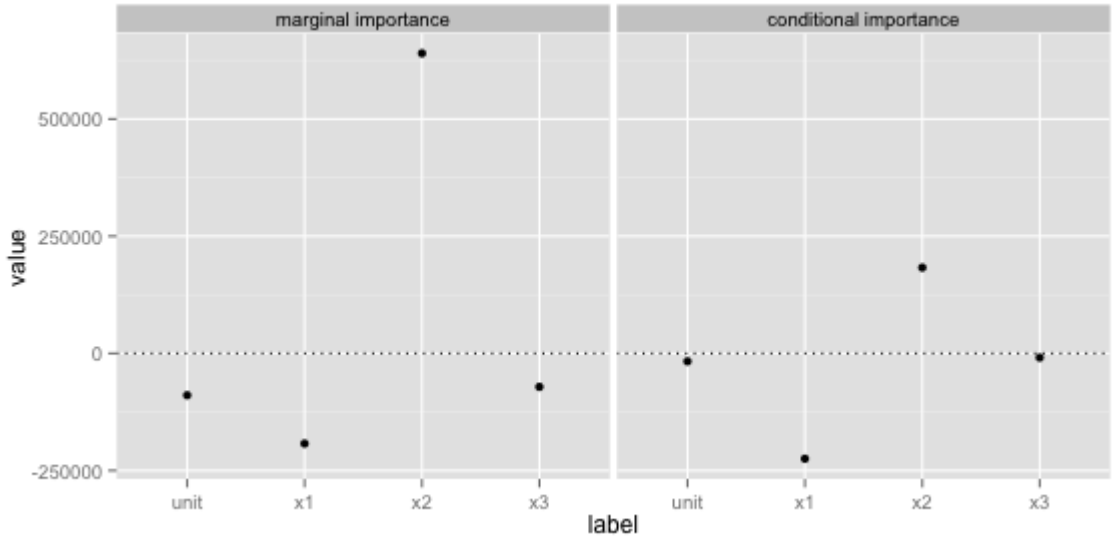

Figure 3: plot of chunk permutation importance

After fitting a random forest and using (marginal/conditional) permutation importance it looks like $\mathbf{x}_2$ is the only variable that is related to $\mathbf{y}$ in this data. These relationships could be spurious though, as with only 25 observations there is a large amount of sampling variablility in these measures. Given that there is a big difference between the marginal and conditional importance for $\mathbf{x}_2$ it seems like there might be some interaction in the regression function.

```
library(edarf)
out <- lapply(c("x1", "x2", "x3", "unit"), function(var) {
    pd <- partial_dependence(fit, df, var, 25)
    colnames(pd)[1] <- "var"
    pd$label <- var
    pd
})
out <- do.call(rbind, out)

ggplot(out, aes(var, pred, group = label)) + geom_line() + geom_point() + facet_grid(~label,
    scales = "free_x")
```

The partial dependence plots seem to agree with the permutation importance. $\mathbf{x}_2$ seems to have a negative and fairly steep relationship with $\mathbf{y}$. Due to the extreme scale of $\mathbf{y}$ and no reference point for what constitutes "important" variation it is unclear whether the relationships found for the other variables are interesting or not. Again, the sampling variation in these relationships might be quite large.

```
out <- partial_dependence(fit, df, c("x3", "x2"), 25)
out$bin[out$x3 >= 75] <- "x3 >= 65 & < 100"
```

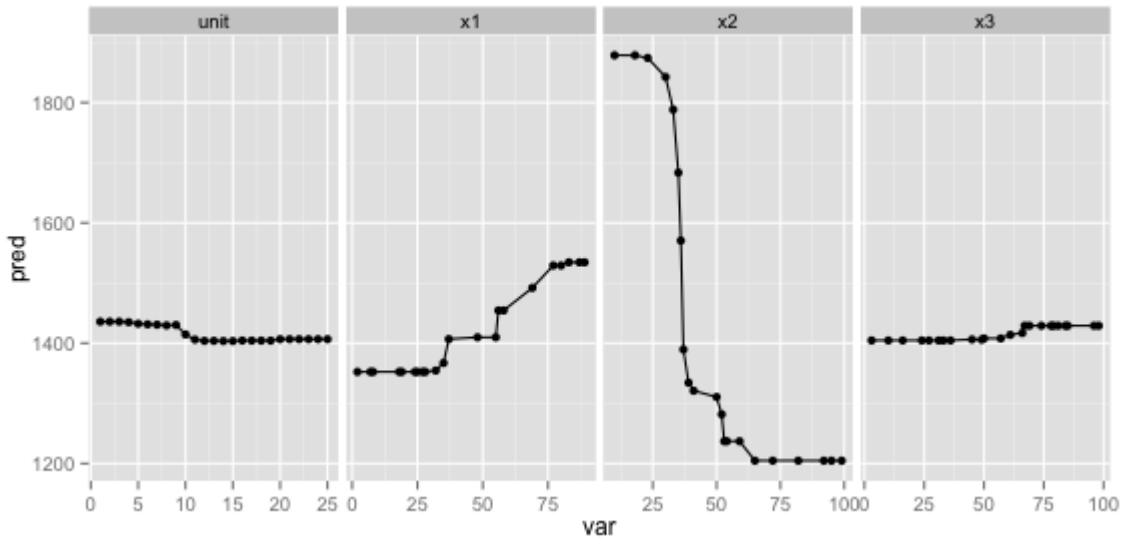

Figure 4: plot of chunk bivariate partial dependence

```
out$bin[out$x3 < 75] <- "x3 < 65 & > 0"
p3 <- ggplot(out, aes(x2, pred)) + geom_point() + geom_smooth(method = "loess",
    se = FALSE) + facet_grid(~bin)

out <- partial_dependence(fit, df, c("x1", "x2"), 25)
out$bin[out$x1 >= 75] <- "x1 >= 75 & < 100"
out$bin[out$x1 < 75 & out$x1 >= 65] <- "x1 >= 65 & < 75"
out$bin[out$x1 < 65 & out$x1 >= 35] <- "x1 >= 35 & < 65"
out$bin[out$x1 < 35] <- "x1 >= 0 & < 35"
p1 <- ggplot(out, aes(x2, pred)) + geom_point() + geom_smooth(method = "loess",
    se = FALSE) + facet_grid(~bin)

out <- partial_dependence(fit, df, c("unit", "x2"), 25)
out$bin[out$unit >= 10] <- "unit >= 10"
out$bin[out$unit < 10] <- "unit < 10"
punit <- ggplot(out, aes(x2, pred)) + geom_point() + geom_smooth(method = "loess",
    se = FALSE) + facet_grid(~bin)

library(gridExtra)
grid.arrange(p1, p3, punit)
```

I don't see much evidence of interaction, though of course our power to discover this type of functional form is probably quite low. I do see again that $\mathbf{y}$ appears to be an increasing function of $\mathbf{x}_1$. I did not investigate 3-way interaction.

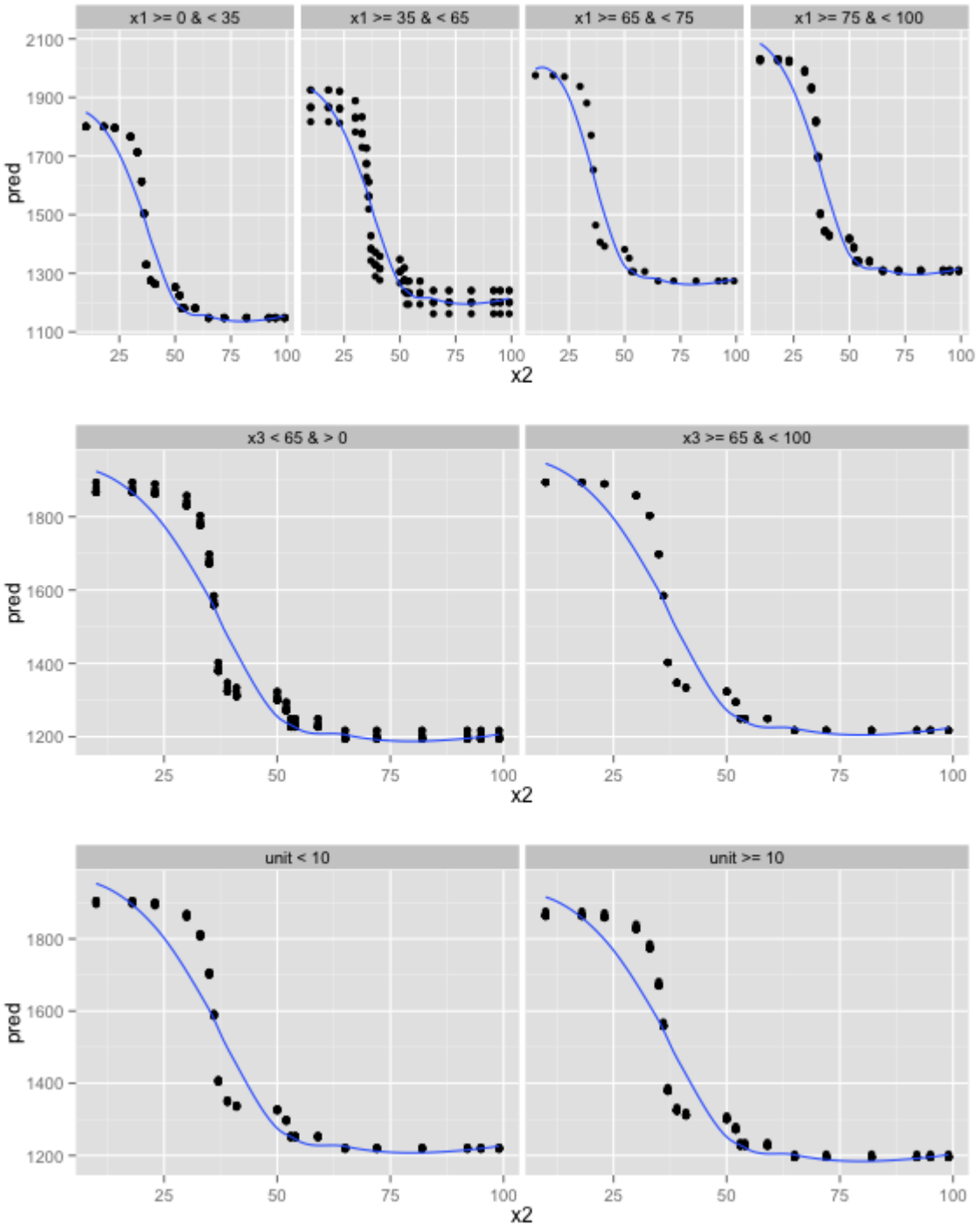

Figure 5: plot of chunk multivariate partial dependence