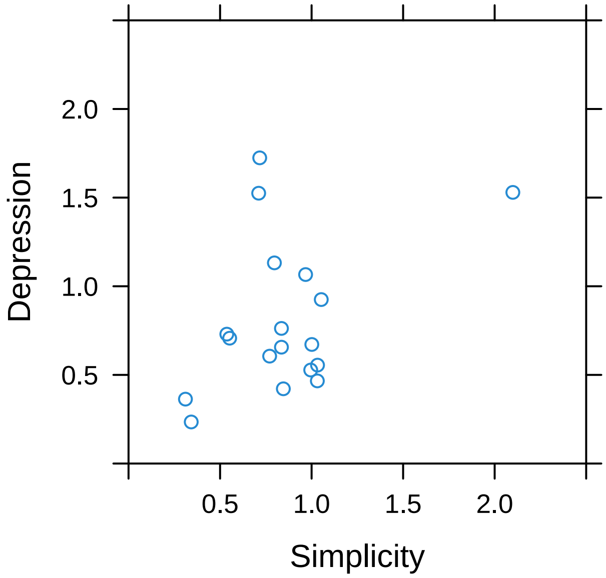
STAT 113: MORE FUN WITH CORRELATION

The dataset `depression.csv` (available at `http://colindawson.net/data/depression.csv`) contains measurements of three quantitative variables for each of 15 patients hospitalized for depression. The variable `Depression` is a standard measure of depression, adjusted to control for some factors known to be correlated with depression. The variable `Simplicity` is a measure of patients' need to see the world in absolutes, or "black and white". A scatterplot relating these two variables is below.



(1) Estimate Pearson's correlation, r, for this sample of patients.

(2) Reproduce the scatterplot, together with a trend line, using the following R code. Estimate the slope of the trend line.

```r
## First load mosaic
library("mosaic")
## Then read in the data
Depression <- read.file("http://colindawson.net/data/depression.csv")
## Make the plot (the xlim and ylim arguments set the range of the axes)
## Reminder: line breaks are just for readability; R doesn't pay attention
## to them
xyplot(Depression ~ Simplicity, data = Depression,
       type = c("p","r"), xlim = c(0,2.5), ylim = c(0,2.5))
```

(3) Compute the actual correlation coefficient. Were you close? (Code is below in case you forget how to do this)

```r
## Correlation
cor(Depression ~ Simplicity, data = Depression)
```

(4) Repeat the previous exercises after `filter()`ing out the single case at the right-hand side of the plot by retaining only those cases that have `Simplicity < 2.0`. How might you characterize the case that you removed?

```r
Depression.filtered <- filter(Depression, Simplicity < 2.0)
xyplot(Depression ~ Simplicity, data = Depression.filtered,
      type = c("p", "r"), xlim = c(0,2.5), ylim = c(0,2.5))
cor(Depression ~ Simplicity, data = Depression.filtered)
```

(5) Comment on the properties of Pearson's correlation and the slope of the line of best fit with respect to the concept of **robustness** (aka, **resistance**).

(6) If you have extra time, repeat the above (plot and correlation with and without the possible outlier), but replacing `Depression` with `rank(Depression)` and similar for `Simplicity`. This converts the actual values to their ranks with respect to the rest of the data: lowest becomes 1, highest becomes n (the sample size), etc. The correlation computed over ranks is called **Spearman's correlation**. How robust/resistant is Spearman's correlation, compared to Pearson's?