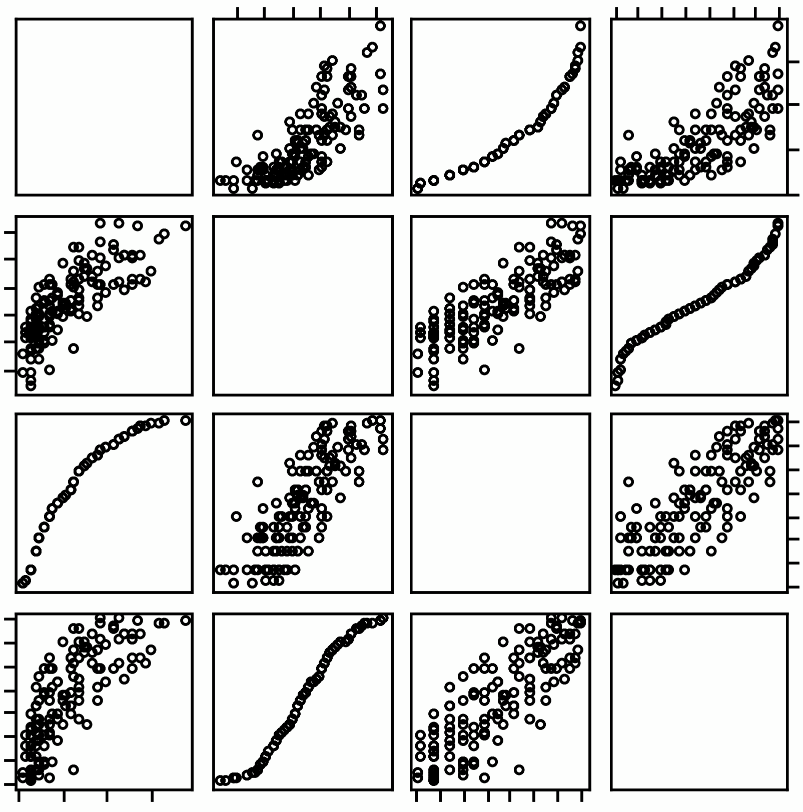
Non-parametric regression

One transformation that get us around many assumptions about distributions, is the rank distribution. In this, one or more variables are replaced by their ranks. The rank transformation simply assigns the value 1 to the smallest observed value, 2 to the next smallest, etc. Of course in practice there are often ties, and some sort of intelligent tie-breaker needs to be employed. Usually this is to assign both values their average rank.

So the data set 74, 75.5, 75.5, 80 becomes 1, 2.5, 2.5, 4.

The R command to do this is, simply, rank(x).

Returning to the ozone data set, lets consider just two variables: ozone and temp. The pairs scatterplot below shows scatterplots featuring these variables and their rank transformations.



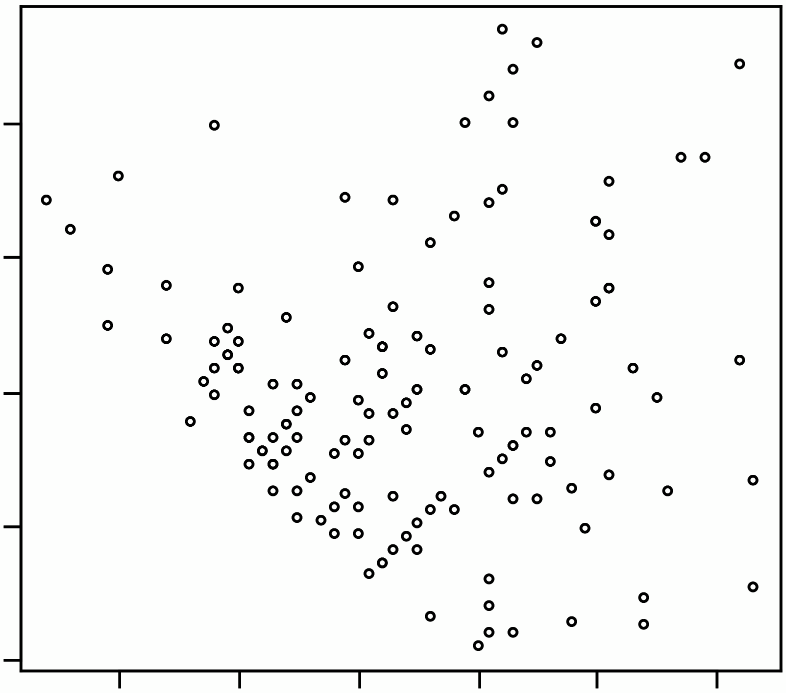
The first box is ozone, next is temp, third is rankozone, and fourth is ranktemp. (For some reason, the wordprocessor will not include the labels of the plot.) Note that the relation between ozone and temp is clearly non-linear. But the ranked transformations are more linear. Here is some R output:

```
> fit2 <- lm(rankozone ~ temp)
> summary(fit2)

Call:
lm(formula = rankozone ~ temp)

Residuals:
    Min      1Q  Median      3Q     Max 
-48.329 -20.184   2.250  16.369  60.210 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -96.8912    10.3153  -9.393 2.22e-16 ***
temp          2.6184     0.1578  16.589  < 2e-16 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1 

Residual standard error: 23.69 on 139 degrees of freedom
Multiple R-Squared: 0.6644,     Adjusted R-squared: 0.662 
F-statistic: 275.2 on 1 and 139 DF,  p-value:     0 

> summary(ozone ~ ranktemp)
 Length   Class    Mode 
      3 formula    call 
> fit3 <- lm(ozone ~ ranktemp)
> summary(fit3)

Call:
lm(formula = ozone ~ ranktemp)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.3025  -3.4997  -0.5813   2.5819  15.7362 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.09802    0.84845   0.116    0.908    
ranktemp     0.16004    0.01037  15.436   <2e-16 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1 

Residual standard error: 5.009 on 139 degrees of freedom
Multiple R-Squared: 0.6316,     Adjusted R-squared: 0.6289 
F-statistic: 238.3 on 1 and 139 DF,  p-value:     0 

> fit4 <- lm(rankozone ~ ranktemp)
> summary(fit4)

Call:
lm(formula = rankozone ~ ranktemp)
```

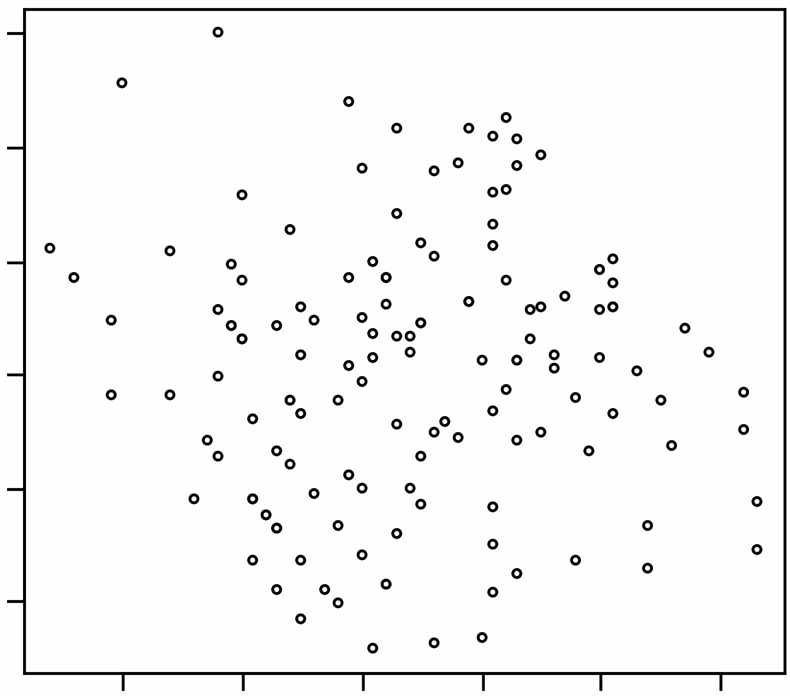
```
Residuals:
    Min      1Q  Median      3Q     Max
-52.016 -16.782   1.018  15.338  67.234

Coefficients:
            Estimate Std. Error t value Pr(>ltl)
(Intercept) 12.25025    3.87536   3.161  0.00193 **
ranktemp     0.82746    0.04736  17.472  < 2e-16 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 22.88 on 139 degrees of freedom
Multiple R-Squared: 0.6871,        Adjusted R-squared: 0.6849
F-statistic: 305.3 on 1 and 139 DF,  p-value:     0
```

The point isn't that we've made tremendous strides in improving the R0squared value, or even that we've learned all that much more. But we've better satisfied the assumptions, as can be seen from looking at residual plots

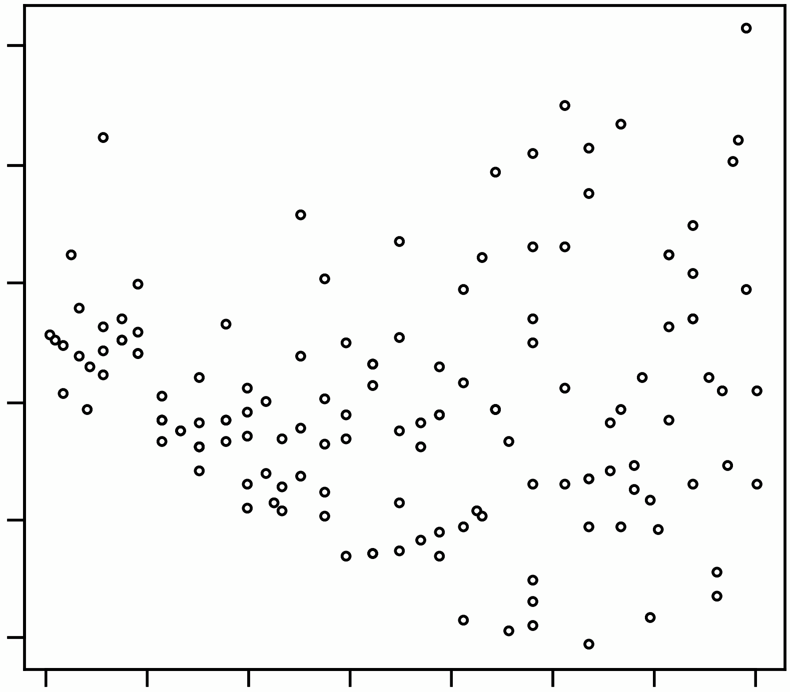
This shows residuals of the first fit (ozone ~ temp) against temp.



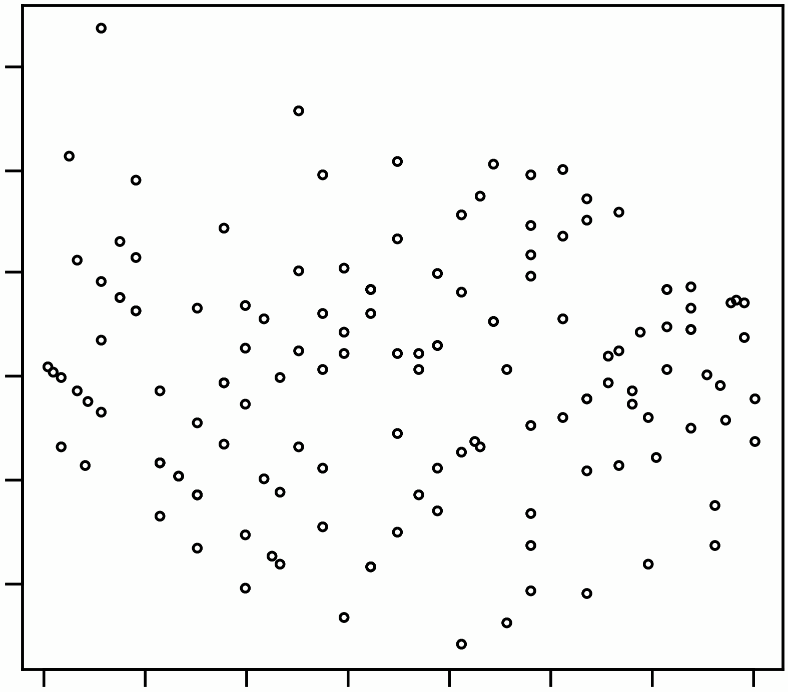
The fit of rankozone against temperature is much better:



The fit with ranking temp, but not ozone, is terrible. Here's the plot of residuals against ranktemp, and had we plotted it against temp it would have been just as bad:



Things improve if we look at the fit in which both variables were transformed. Here's the residuals of that fit against ranktemp:



Interpreting these models is more difficult. In the first, untransformed, fit we learned that the slope between ozone and temperature was statistically significant, which means we could be confident that it is non-zero and there is a "real" relation between ozone and temp. However, this conclusion holds only to the extent that the assumptions are satisfied, and even a casual inspection of the plot shows that the relation is not linear.

The transformed data better satisfy the assumptions, and so we are more certain of our conclusion that the slope is significant. What we've lost is the ability to interpret the numerical value of the slope. The estimated value of .82746 means that when we increase to the next largest temperature, the rank of ozone goes up, on average, by .828. But this doesn't really translate well into a scientific model. What does translate is that the increasing relationship is 'real".