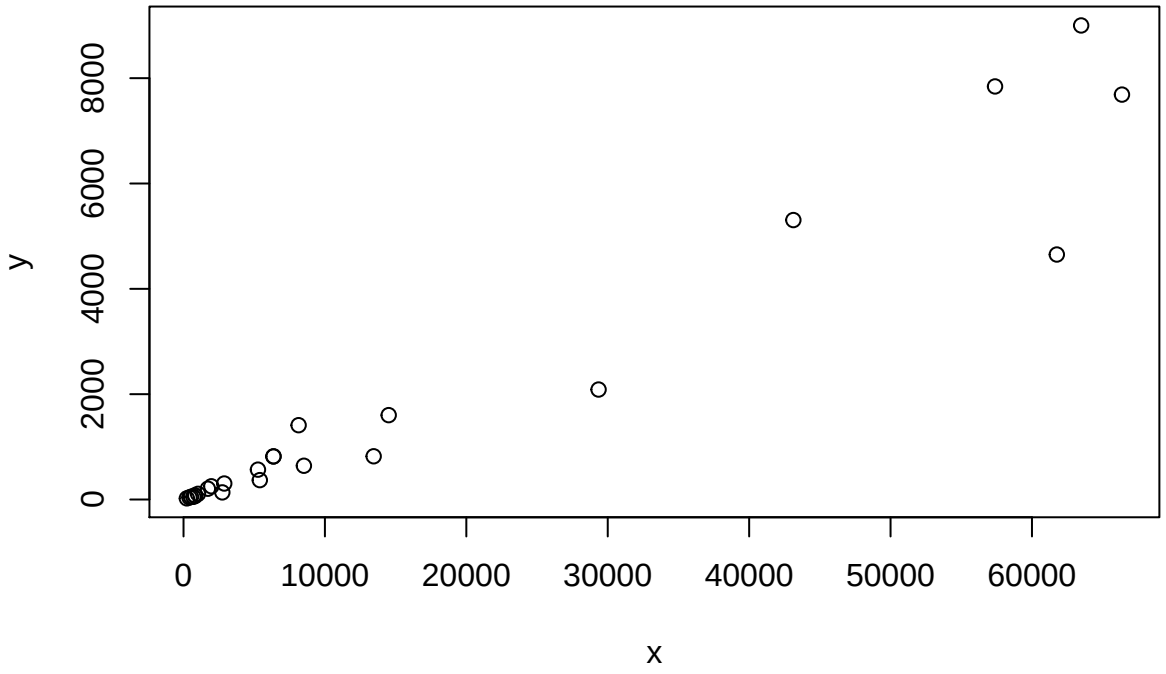
```r
# Read data from florange.csv and input it into the dat vector.
dat <- read.csv("florange.csv")
# Done to make the predict function work well.
x <- dat$acres
y <- dat$boxes
# Output the first 6 rows in dat.
head(dat)
```

```
##      county boxes acres
## 1   Brevard    51   696
## 2 Charlotte   821 13447
## 3   Collier  2088 29351
## 4    DeSoto  7688 66365
## 5    Glades   368  5396
## 6    Hardee  5306 43126
```

```r
# Draw a scatterplot with x-axis as `acres` and y-axis as `boxes`.
plot(x,y)
```




```r
# Compute some common variables with common functions.
r <- cor(x,y)
xbar <- mean(x)
ybar <- mean(y)
cat("r:", r, "xbar:", xbar, "ybar:", ybar)
```

```
## r: 0.9635098 xbar: 16132.64 ybar: 1797.56
```

Therefore, $r = 0.9635098$, $\bar{x} = 16132.64$, and $\bar{y} = 1797.56$.

```r
# Compute some common variables manually.
Sxx <- sum( (x - xbar)^2 )
Sxy <- sum( (x - xbar) * (y - ybar) )
cat("Sxx: ", Sxx, "Sxy: ", Sxy)
```

```
## Sxx:  12450023404 Sxy:  1453128337
```

Therefore, $S_{xx} = 12450023404 = 1.245 \times 10^{10}$ and $Sxy = 1453128337 = 1.453 \times 10^{9}$.

```
# R's lm function fits linear models
lm.1 <- lm(y~x)
summary(lm.1)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2470.81    -6.17    71.72   106.46  1677.32
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -85.391989 186.178031  -0.459    0.651
## x             0.116717   0.006761  17.263 1.16e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 754.4 on 23 degrees of freedom
## Multiple R-squared:  0.9284, Adjusted R-squared:  0.9252
## F-statistic:   298 on 1 and 23 DF,  p-value: 1.164e-14
```

From the summary, we can see that $\hat{\beta}_0 = -85.391989$, $\hat{\beta}_1 = 0.116717$, $\text{Se}(\hat{\beta}_1) = 0.006761$, $t = 17.263$, p-value $= 1.64 \times 10^{-14}$, and $\hat{\sigma} = 754.4$.

```
# Sum Squared Fitted Values
sum(lm.1$fitted.values^2)
```

```
## [1] 250385207
```

```
# Sum Squared Residuals
sum(lm.1$residuals^2)
```

```
## [1] 13089860
```

Therefore, $SS(\text{Res}) = \sum_{i=1}^{n} e_i^2 = 13089860 = 1.31 \times 10^{7}$.

```
# Manual calculation of sigma^2 estimate
sum(lm.1$residuals^2) / 23
```

```
## [1] 569124.3
```

Therefore, $\hat{\sigma}^2 = 69124.3 = 5.7 \times 10^{5}$.

```
# Manual calculation of sigma estimate
sqrt(sum(lm.1$residuals^2) / 23)
```

```
## [1] 754.4033
```

Therefore, $\hat{\sigma} = 754.4$.

```
# t distribution values
qt(0.975,23)
```

```
## [1] 2.068658
```

Therefore, $c = 2.07$.

```r
# 95% confidence interval
confint(lm.1)
```

```
##                    2.5 %      97.5 %
## (Intercept) -470.5305905 299.7466119
## x              0.1027305   0.1307034
```

```r
# 95% prediction interval with predicted boxes if we had 10000 acres
predict(lm.1, data.frame(x=10000), interval="prediction")
```

```
##        fit       lwr      upr
## 1 1081.777 -512.0407 2675.595
```

Q: Is σ the same for all values of y?

A: It appears to not in the sense that the variance appears to be higher with respect to higher acres. Sigma will be smaller when there's less acres. Later, this will be testing equal variance or homoscedastic assumption. Later, when we talk about variable transformations we can consider taking the logarithm.

Q: Are the error terms plausibly independent? In other words, does knowing one e_i (residual) help predict e_j (another residual) for a different county?

A: There's diagnostics for checking this. However, intuitively there could be some common factors at play when two counties are geographically close.