# The Simplest Introduction to Regression EVER

*J. Mao*

We observe $x$ and $y$ in the data.

Suppose we know that $x$ and $y$ have a linear relationship, like this:

$$y = a + b * x + e$$

, where e is random noise.

We know $x$ and $y$ have this kind of relationship, but we do not know the values of $a$ and $b$. What can we do?

Use regression! Regression is a method that helps us to find the values of $a$ and $b$ using data on $x$ and $y$.

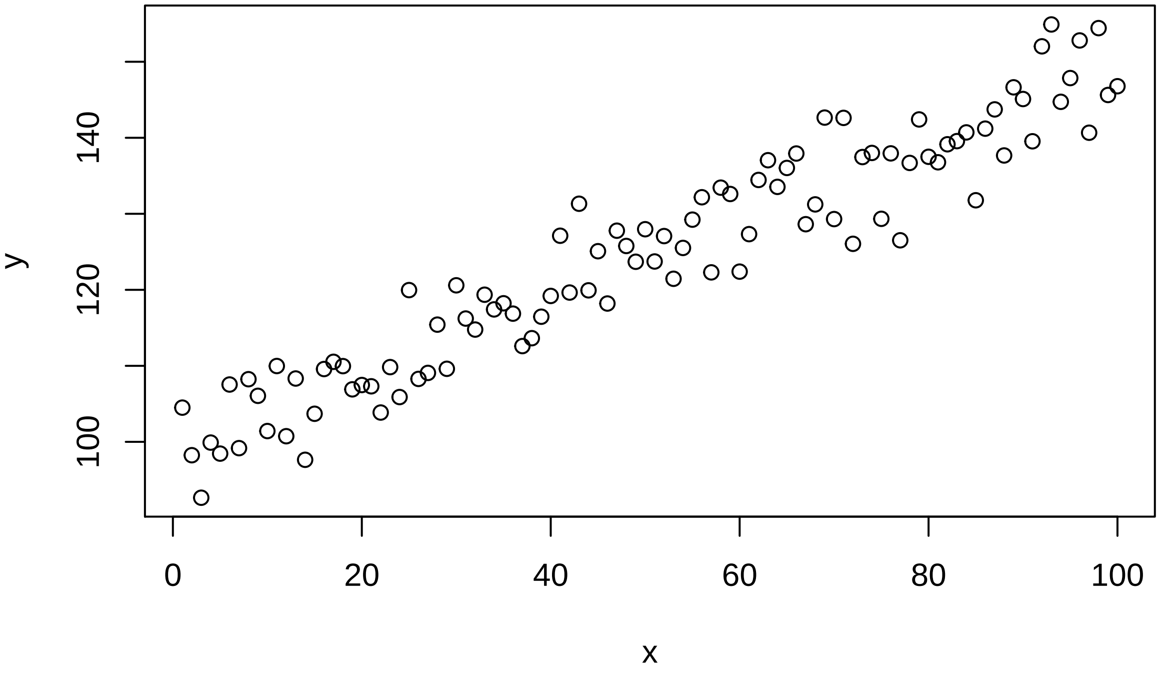
Here's how to do regression in R:

Firt, let's generate some data:

```
x = 1:100
e = rnorm(length(x),mean=0,sd=5) #generate e from normal distribution with mean 0 and standard deviatio
y = 100 + 0.5*x + e
```

Let's plot the data:

```
plot(x,y)
```



Since we generated this data set, we know that $y = a + b * x + e$, where $a = 100$ and $b = 0.5$. But suppose we do *not* know and would like to find out the value of $a$ and $b$ from our data, here is how we do it:

```
lm(y~x) #regression of y on x
```

```
##
## Call:
## lm(formula = y ~ x)
##
```

```
## Coefficients:
## (Intercept)            x
##     99.1089       0.5021
```

That's it! To find out the value of $a$ and $b$, we perform a regression of $y$ on $x$ (or we say "regress $y$ on $x$"), which in `R`, is simply `lm(y~x)`.

From the output of `lm(y~x)`, look at "Coefficients". It tells you that the "Intercept" is about 100 and "x" is about 0.5. Here the "Intercept" is the $a$ in our model and "x" is the $b$ in our model and you can see that the estimates are pretty close to their real value!

You may recall that when we draw a best linear fit line onto a scatter plot, the command is: `abline(lm(y~x))`. Now you should understand that this is basically a command that tells `R` to draw the linear regression line!