## Using R to perform simple linear regression

In statistics, modelling quantifies the relationships between our data variables in hand and from the model obtained, we can make predictions on the possible outcome.

We know a simple linear regression is the most basic model with two variables, $x_i$ and $y_i$, and is modeled as a linear relationship with an error term $\varepsilon_i$ as shown below:

$$y_i = a + bx_i + \varepsilon_i$$

where coefficients $a$ is the $y$-intercept and $b$, the gradient of the straight line curve. So, our task is to find a way to use these variables ($x$ being the *predictors* and $y$, the *response*) in fitting into a linear model in order to get the best estimates for $a$ and $b$. In a laboratory instrumental calibration work, we usually have $x$ as a series of concentrations of prepared working standard solutions and $y$, the instrument's responses.

The beauty of R is that it can be used easily to build up a linear regression model, calculate its gradient, $b$ and $y$-intercept, $a$, and provide a full regression summary of the statistical results from the plot. At the same time, we can also answer the following questions:

- *Is the linear model statistically significant?*
  We look at the $F$ statistic given at the bottom of this summary
- *Are the coefficients significant?*
  Check the coefficient's $t$-statistics and $p$-values in the summary
- *Is the model useful?*
  Check the $R^2$ near the bottom of the summary

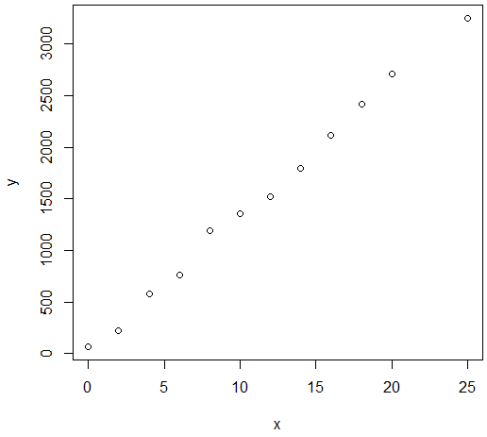
Let's see how the R language is used to perform a simple linear regression.

Assume we have obtained a series of analyte concentrations of working standards against the instrumental responses as below:

```
x<- c(0,2,4,6,8,10,12,14,16,18,20,25)
y<- c(70,220,580,758,1194,1360,1524,1800,2112,2418,2710,3250)
```

The "plot(x,y)" function will plot these pairs of data points:

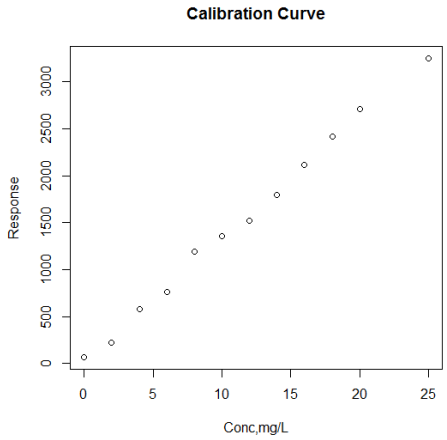
```
> x<-c(0,2,4,6,8,10,12,14,16,18,20,25)
> y<-c(70,220,580,758,1194,1360,1524,1800,2112,2418,2710,3250)
> plot(x,y)
```




We can label the plot by adding in more commands:

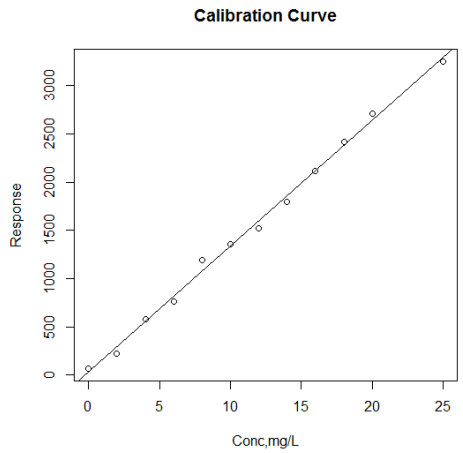
```
plot(x,y, xlab="Conc,mg/L", ylab="Response", main="Calibration Curve")
```

and hence we get:




If we wish to add a line to the plot to illustrate the linear regression between the $x$'s and $y$'s, we create a model object, plot the $(x,y)$ pairs, and then plot the model object using the "lm(y~x)" and "abline( )" functions

```
> x<- c(0,2,4,6,8,10,12,14,16,18,20,25)
> y<- c(70,220,580,758,1194,1360,1524,1800,2112,2418,2710,3250)
> lm(y~x)
> plot(y~x,xlab="Conc,mg/L",ylab="Response",main="Calibration Curve")
> abline(lm(y~x))
```




To quantify the above regression model of the relationship, we just simply type :

```
> lm(y~x)
```

and the output is:

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

Coefficients:
(Intercept)            x
      32.88       130.38
```

In this case, the regression equation is:

$$y_i = 32.88 + 130.38x_i$$

To obtain the other critical regression statistics, we use the "summary" function:

```
> summary(lm(y~x))
```

and hence, the outputs are:

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

Residuals:
    Min      1Q  Median      3Q     Max
-73.638 -57.427   8.166  37.408 118.073

Coefficients:
```

```
             Estimate     Std. Error   t value    Pr(>|t|)
(Intercept)   32.876       34.297     0.959       0.36
x            130.381        2.553    51.062    2.01e-13 ***
---
Signif. codes:   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 64.91 on 10 degrees of freedom
Multiple R-squared:   0.9962,      Adjusted R-squared:   0.9958
F-statistic:   2607 on 1 and 10 DF, p-value: 2.007e-13
```

From the summary of statistical data, we note that:

1. The F-statistic tells us whether the above model is significant or not. The model is significant if any of the coefficients are non-zero. Conventionally, a $p$-value of <0.05 indicates that the model is likely significant (i.e. one or more coefficients are non-zero). The reverse is true. In here, we see $F$=2607 (deg of freedoms: 1, 10) with $p$-value of much less than 0.0001. That means the probability is only 0.0001 that our model is insignificant!

2. The adjusted *coefficient of determination*, $R^2$ value of 0.9958 indicates that in this regression model has 99.58% of the variance of $y$ and the remaining 0.42% is due to other unexplained factors. As we know, bigger $R^2$ is better for the model's quality.

3. In the $t$-statistic evaluation, we have the following results:

```
Coefficients:
             Estimate     Std. Error   t value    Pr(>|t|)
(Intercept)   32.876       34.297     0.959       0.36
x            130.381        2.553    51.062    2.01e-13 ***
---
Signif. codes:   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What does that whole lot of figures mean to us?

First, the columns labelled **Estimate** contains the estimated regression coefficients (i.e. $a$ and $b$) as calculated by the ordinary least squares (OLS) method.

We would then ask on statistically speaking, how likely is it that these coefficients are truly zero. The $t$-statistic and the $p$-values in the summary answer our question.

Since the p-value is a probability, it gauges the likelihood that the coefficient is not significant, so smaller is better. A big $p$-value is bad because it indicates a high likelihood of insignificance. In this example, our p-value for the gradient, b, is a mere $2.01 \times 10^{-13}$ and hence, we conclude that the gradient $b$ is very likely significant.

Furthermore, we see another line in the summary stating a series of significant codes with asterisks (*) and also at the *last* column. This is a handy feature of R to flag the significant variables for our quick identification. The line labelled "Signif. codes" at the bottom gives a cryptic guide to the flags' meanings:

| | |
|---|---|
| *** | p-value between 0 and 0.001 |
| ** | p-value between 0.001 and 0.01 |
| * | p-value between 0.01 and 0.05 |
| . | p-value between 0.05 to 0.1 |
| (blank) | p-value between 0.1 and 1.0 |

The other parameters to be noted are:

a. The column labelled "std. Error" is the standard error of the estimated coefficient.
b. The column labelled "$t$-value" is the t-statistic from which the $p$-value was calculated.