Simple Linear Regression Model Example using NASA global average surface temperature by year data

NASA's Goddard Institute for Space Studies records the average surface temperature across the globe each year and compares this to a baseline of the average temperature for the 30 year period 1951 – 1980. The data is available at https://climate.nasa.gov/vital-signs/global-temperature/ and data for the years since 1980 is saved in a .csv file on QM Plus. The full dataset online goes back to 1880.

We can use this data to construct a Simple Linear Regression Model in R which we will use in the lectures.

First we import the data from the .csv file into our R workspace

```
> Global_Temperature_NASA_Data <- read.csv("~/5120 Statistical Modelling I 2022-23/Global_Temperature_NASA_Data.csv")
```

```
>   View(Global_Temperature_NASA_Data)
```

We can assign the Year and Temperature columns in the dataframe to variables *x* and *y* for our analysis. We will work with *x* = Year – 1980 rather than the full year so that the explanatory variable is years elapsed since the baseline period ended.

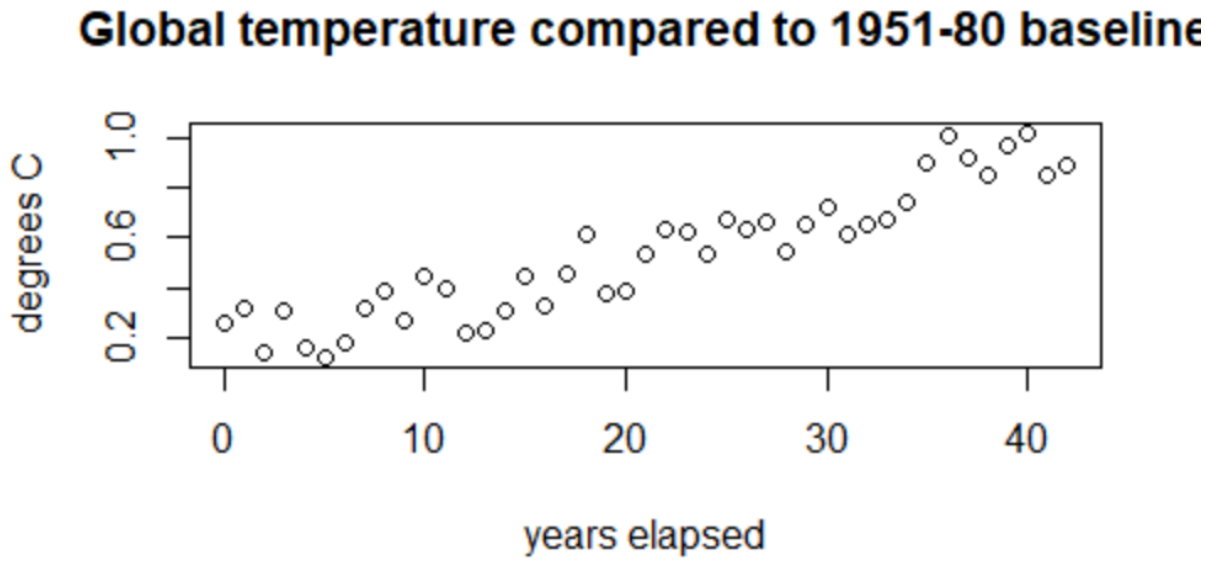
Note that strictly in R we do not need to assign the data to vectors x and y. The R functions will work by calling the relevant columns in the dataframe `Global_Temperature_NASA_Data$Year` *and* `Global_Temperature_NASA_Data$Temperature` *instead. However to keep the notation similar to the x and y in the lecture material, we assign to vectors here.*

```
> x <- Global_Temperature_NASA_Data$Year - 1980
```

```
> y <- Global_Temperature_NASA_Data$Temperature
```

Now we can construct a scatterplot.

```
> plot(x,y, main = "Global average surface temperature compared to 1951-80 baseline", xlab = "years elapsed", ylab = "degrees C")
```



The initial view of this scatterplot suggests that a linear regression model for global annual average surface temperature by years elapsed since the baseline measurement may well be appropriate.

We can use the 'linear model' `lm()` function in R to find the least squared estimates of the simple linear regression model parameters

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

Call:
lm(formula = y ~ x)

Residuals:
      Min        1Q    Median        3Q       Max
-0.153268 -0.080700 -0.003953  0.080943  0.193511

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.138404   0.028516   4.854 1.79e-05
x           0.018836   0.001169  16.112  < 2e-16
```

```
(Intercept) ***
x           ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09513 on 41 degrees of freedom
Multiple R-squared:  0.8636,  Adjusted R-squared:  0.8603
F-statistic: 259.6 on 1 and 41 DF,  p-value: < 2.2e-16
```

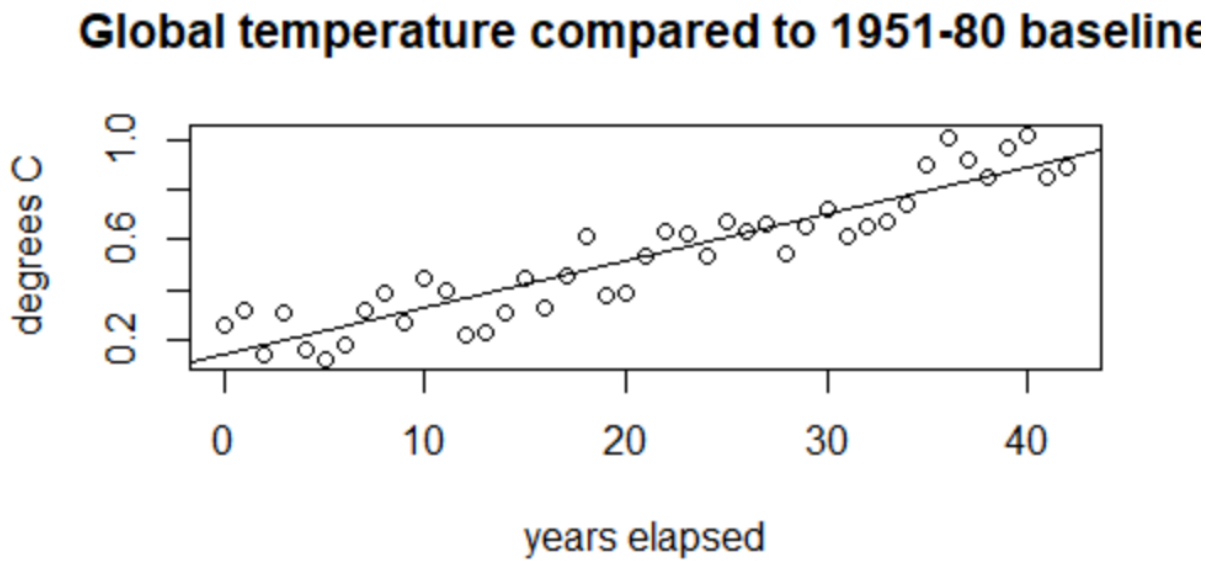
There is a lot of information on the regression model in the `summary()` function which we will explore in later weeks on the module, but for now we can use `lm()` to give us the least squares estimates which we denoted $\widehat{\beta_0}$ and $\widehat{\beta_1}$ in the lecture notes and are presented as the `Estimate` of the `(Intercept)` and the `x Coeffients` here in the R output.

Intercept coefficient estimate ($\widehat{\beta_0}$) = 0.138404

Slope coefficient estimate ($\widehat{\beta_1}$) = 0.018836

We can add the fitted regression line to the earlier plot

```
> abline(model)
```



Interpretation of the model result

The simple linear regression model fitted may be interpreted as:

- by the end of the 1951 – 1980 baseline period, global average surface temperatures were already 0.138 degrees higher than the average for the 30-year baseline period
- since then annual average surface temperatures have increased by 0.0188 degrees Celsius per year on average based on this NASA GISS data.