OLS coefficients by hand in R and Python

Deriving the OLS coefficients

The independent and depenendent variables in a multivariate regression can be represented in matrix notation as

$$y = X\beta + u,$$

where

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{T1} & x_{T2} & \cdots & x_{Tk} \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix}, \quad u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_T \end{pmatrix}.$$

In matrix notation, the criterion function to be minimized is

$$SSE(\beta) = (y - X\beta)'(y - X\beta),$$

and the first-order conditions are

$$\frac{\partial SSE(\beta)}{\partial \beta} = -2X'(y - X\hat{\beta}) = 0,$$

which yields the normal equations,

$$(X'X)\hat{\beta} = X'y.$$

As long as $(X'X)$ is of full rank, then

$$\hat{\beta} = (X'X)^{-1}X'y.(\#eq : beta) \tag{1}$$

It can be shown via the Gauss-Markov theorem that under the classical assumptions, the OLS estimator has the least variance in the class of all linear unbiased estimators of β. However, the point of this document is to show how to calculate the OLS coefficients by hand using the computer programs R and Python. Let's start with R.

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

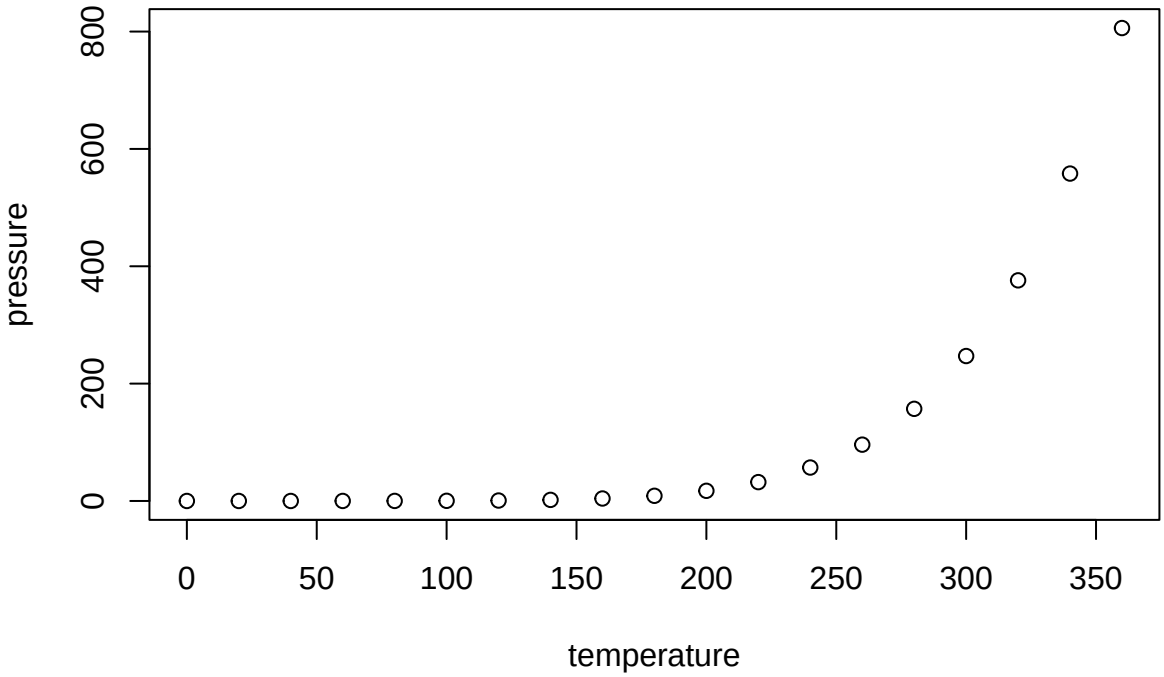
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.