R Tutorial: Intro to Simple Linear Regression

For this tutorial, we will use data that you saw in the readings for the module, the rocket propellant data. The textbook provides the description of the data as:

A rocket motor is manufactured by bonding an igniter propellant and a sustainer propellant together inside a metal housing. The shear strength of the bond between the two types of propellant is an important quality characteristic. It is suspected that shear strength is related to the age in weeks of the batch of sustainer propellant.

Download the data file, `rocket.csv`, from Collab and read the data in. Also load the `tidyverse` package.

```
library(tidyverse)
```

```
Data<-read.csv("rocket.csv", header=TRUE)
head(Data)
```

```
##   Observation..i Shear.Strength..yi..psi. Age.of.Propellant..xi..weeks.
## 1              1                  2158.70                         15.50
## 2              2                  1678.15                         23.75
## 3              3                  2316.00                          8.00
## 4              4                  2061.30                         17.00
## 5              5                  2207.50                          5.50
## 6              6                  1708.30                         19.00
```

Notice there is an extra column for the observation number, and the names of the columns are long and complicated. So we remove the first column and rename the 2nd and 3rd columns

```
##remove first column
Data<-Data[,-1]
##rename the remaining 2 columns
names(Data)<-c("Strength", "Age")
head(Data)
```

```
##   Strength   Age
## 1  2158.70 15.50
```

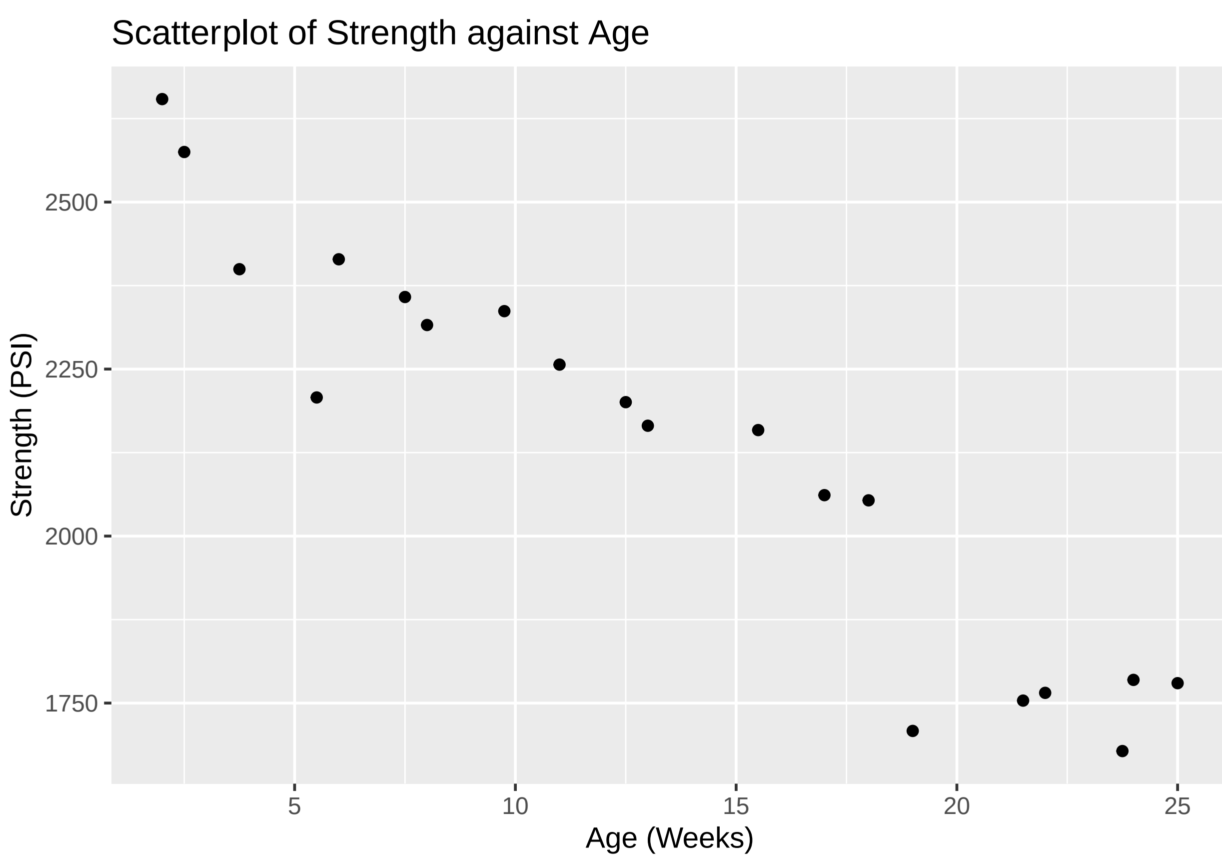
```
## 2  1678.15 23.75
## 3  2316.00  8.00
## 4  2061.30 17.00
## 5  2207.50  5.50
## 6  1708.30 19.00
```

The data frame looks a lot neater now.

1. Scatterplots

One of the first things to do is to create some data visualizations of our data. Since we have two quantitative variables, a scatterplot should be used. To create a scatterplot of `Strength` against `Age`

```
ggplot(Data, aes(x=Age,y=Strength))+
  geom_point()+
  labs(x="Age (Weeks)", y="Strength (PSI)",
       title="Scatterplot of Strength against Age")
```



or, using base R functions

```
plot(Data$Age, Data$Strength,
     xlab="Age (Weeks)",
     ylab="Strength (PSI)",
     main="Scatterplot of Strength against Age")
```

Based on the scatterplot, we see that the strength of the bond has a negative linear relationship with the age of the propellant, i.e. as the propellant gets older, the strength of the bond weakens.

Note: the `ggplot()` function is the only function in this tutorial that is not a base R function.

2. Saving plots as an external file

If you want to save your plot to an external file

```
jpeg("plot.jpg")
ggplot(Data, aes(x=Age,y=Strength))+
  geom_point()+
  labs(x="Age (Weeks)",
       y="Strength (PSI)",
       title="Scatterplot of Strength against Age")
dev.off()
```

A line of code is placed before creating the plot, and a line `dev.off()` is placed after the plot. The plot is then saved in the working directory. If you want to save the plot as a pdf, use `pdf("plot.pdf")` instead for the first line. You will notice that R does not produce the plot in the plot window.

Alternatively, you can click on `Plots`, and then `Save as Image...` or `Save as PDF...` in RStudio.

3. Simple Linear Regression

We use the `lm()` function to fit a simple linear (SLR) model to our data

```
result<-lm(Strength~Age, data=Data)
```

Inside `lm()`, we specify the response variable, then the predictor, with a `~` in between, and also specify the name of the data frame via `data`. To view some information stored in `result`

```r
summary(result)
```

```
##
## Call:
## lm(formula = Strength ~ Age, data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -215.98  -50.68   28.74   66.61  106.76
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2627.822     44.184   59.48  < 2e-16 ***
## Age          -37.154      2.889  -12.86 1.64e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 96.11 on 18 degrees of freedom
## Multiple R-squared:  0.9018, Adjusted R-squared:  0.8964
## F-statistic: 165.4 on 1 and 18 DF,  p-value: 1.643e-10
```

We see the following values:

- $\hat{\beta}_1 = -37.154$
- $\hat{\beta}_0 = 2627.822$
- $R^2 = 0.9018$
- $s = 96.11$

So the estimated regression equation is $\hat{y} = 2627.822 - 37.154x$.

To see other pieces of information that can be extracted from `result`

```r
names(result)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

To extract specific information, for example, the vector of residuals

```r
result$residuals
```

```
##          1          2          3          4          5          6
##  106.758301  -67.274574  -14.593631   65.088687 -215.977609 -213.604131
##          7          8          9         10         11         12
##   48.563824   40.061618    8.729573   37.567141   20.374323  -88.946393
##         13         14         15         16         17         18
##   80.817415   71.175153  -45.143358   94.442278    9.499187   37.097528
##         19         20
##  100.684823  -75.320154
```

4. ANOVA table

To obtain the corresponding ANOVA table

```r
anova.tab<-anova(result)
anova.tab
```

```
## Analysis of Variance Table
##
## Response: Strength
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Age        1 1527483 1527483  165.38 1.643e-10 ***
## Residuals 18  166255    9236
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The first line of the output gives SS_R, the second line gives SS_{res}. The function doesn't provide SS_T, but we know that $SS_T = SS_R + SS_{res}$.

Again, to see what can be extracted from `anova.tab`

```r
names(anova.tab)
```

```
## [1] "Df"      "Sum Sq"  "Mean Sq" "F value" "Pr(>F)"
```

So SS_T can be easily calculated

```r
SST<-sum(anova.tab$"Sum Sq")
SST
```

```
## [1] 1693738
```

The R^2 was reported to be 0.9018. To verify

```r
anova.tab$"Sum Sq"[1]/SST
```

```
## [1] 0.9018414
```