Multiple Regression in R

Today we will talk about how to run a multiple linear regression in `R`.

As an example, we have data from a hypothetical company's advertising spending and sales for the last three years. The variables are `TV`, `Radio`, and `Newspaper` which are the amount spent on the respective ads in thousands of dollars. The final variable is `Sales` which is that month's sales in units.

Download the data and import it to RStudio. (https://raw.githubusercontent.com/moravian-mspa/MGMT555/master/m5_Tutorial_MLR_in_R/Advertising.csv)

```
library(tidyverse)
Advertising <- read_csv("Advertising.csv");
```

```
Parsed with column specification:
cols(
  Month = col_double(),
  TV = col_double(),
  Radio = col_double(),
  Newspaper = col_double(),
  Sales = col_double()
)
```

```
attach(Advertising)
head(Advertising)
```

```
# A tibble: 6 x 5
  Month    TV Radio Newspaper Sales
  <dbl> <dbl> <dbl>     <dbl> <dbl>
1     1 230.   37.8      69.2   663
2     2  44.5  39.3      45.1   312
3     3  17.2  45.9      69.3   279
4     4 152.   41.3      58.5   555
5     5 181.   10.8      58.4   387
6     6   8.7  48.9      75     216
```

We should look at some summary statistics briefly to ensure everything seems reasonable.

```
summary(Advertising)
```

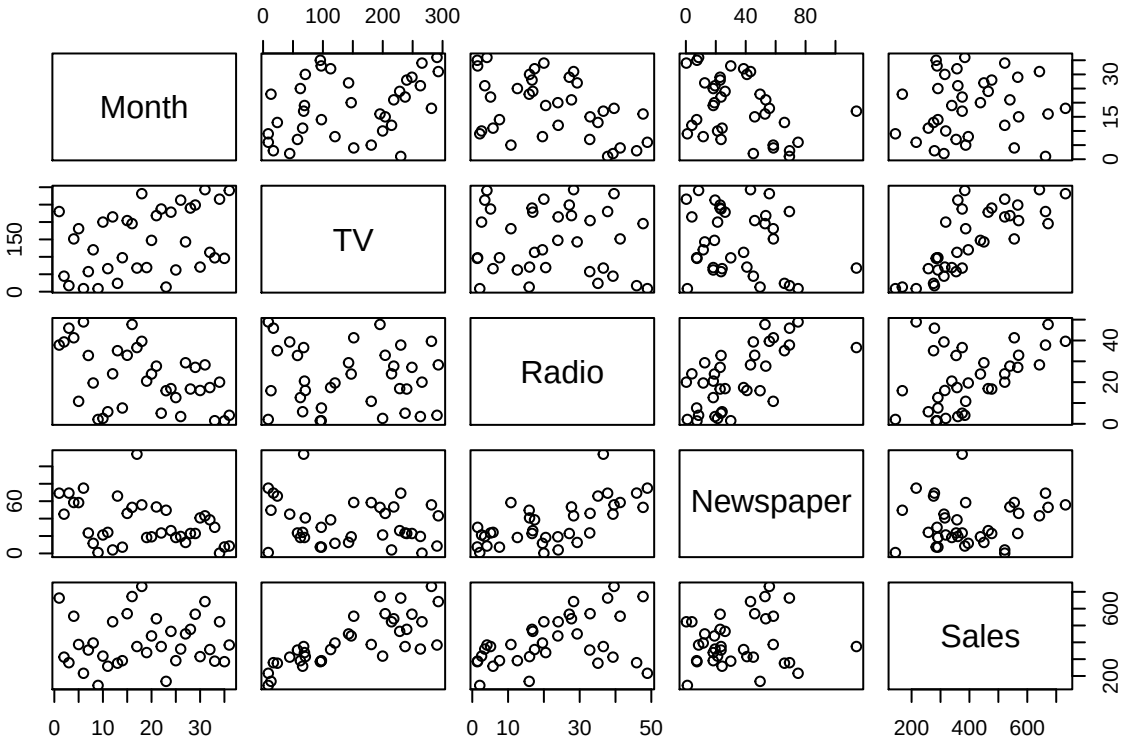
```
     Month             TV             Radio         Newspaper
 Min.   : 1.00   Min.   :  8.60   Min.   : 1.40   Min.   :  0.30
 1st Qu.: 9.75   1st Qu.: 67.38   1st Qu.:10.00   1st Qu.: 18.30
 Median :18.50   Median :145.10   Median :20.25   Median : 25.20
 Mean   :18.50   Mean   :146.28   Mean   :22.18   Mean   : 34.94
 3rd Qu.:27.25   3rd Qu.:228.75   3rd Qu.:33.45   3rd Qu.: 53.02
 Max.   :36.00   Max.   :292.90   Max.   :48.90   Max.   :114.00
     Sales
 Min.   :144.0
 1st Qu.:291.0
 Median :375.0
 Mean   :405.1
 3rd Qu.:522.0
 Max.   :732.0
```

There do not seem to be any unusual values, though, the spread in the TV column is notably large. This is

nothing to worry about just yet, but an interesting component in the dataset.

Let's check to see if the relationships are roughly linear. This command generates dotplots for each possible combination of the variables.

```
pairs(Advertising)
```



Note we will ignore the `Month` row and column for now. If we look at the `Sales` row (the bottom row) we see that `TV` and `Radio` seem to have roughly linear relationships, but `Newspaper` doesn't show much of a pattern.

Well let's try an initial model with all three predictor variables and see what it looks like. The process for creating the model is very similar to simple linear regression from before. Notice we put `Sales` first since that is the dependent variable. Next we add all our independent variables (also called *factors* or *predictor variables*). This is telling `R` that we would like a model of the form:

$\texttt{Sales} = b_0 + b_1\texttt{TV} + b_2\texttt{Newspaper} + b_3\texttt{Radio}$

The software will attempt to find the coefficients to minimize the squared residuals similar to the process for simple linear regression.

```
model <- lm(Sales ~ TV + Newspaper + Radio);
summary(model)
```

```
Call:
lm(formula = Sales ~ TV + Newspaper + Radio)

Residuals:
    Min      1Q  Median      3Q     Max
-147.90  -27.41   16.41   37.81   74.63

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 104.9599    24.7455   4.242 0.000177 ***
```

```
TV           1.2859     0.1038  12.390 9.39e-14 ***
Newspaper   -0.3562     0.4958  -0.718 0.477686
Radio        5.6130     0.8516   6.591 1.98e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 55.14 on 32 degrees of freedom
Multiple R-squared:  0.8687,    Adjusted R-squared:  0.8564
F-statistic: 70.56 on 3 and 32 DF,  p-value: 3.38e-14
```

The very last line gives us the overall significance of the model. It is the result of an ANOVA test which helps to decide if we have eveidence that any of the coefficients are non-zero. Since the p-value is very small, we have strong evidence that at least one of the coefficients should be non-zero. Thus, we have strong evidence there is a linear relationship present in the data.

Directly above that we see the R^2 and adjusted R^2 values. We will use the adjusted R^2 because we are using multiple predictor variables. An adjustmented R^2 of 0.8956 is quite good, and means approximately 90% of the variation in `Sales` is explained by the linear relationship to the three factors.

Consider the "Coefficients" section. There is a separate t-test for each coefficient to see how significant it is. The `Newspaper` variable is the only one whose p-value is too large. We do not have strong evidence that the `Newspaper` coefficient should be non-zero.

Prehaps we should create a model without `Newspaper`. Notice the code below is very similar to the call above, it just omits `Newspaper`:

```
model2 <- lm(Sales~TV+Radio, data=Advertising);
summary(model2)
```

```
Call:
lm(formula = Sales ~ TV + Radio, data = Advertising)

Residuals:
    Min      1Q  Median      3Q     Max
-149.53  -28.00   13.35   41.02   69.98

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  99.2844    23.2781   4.265 0.000158 ***
TV            1.3002     0.1011  12.859 2.11e-14 ***
Radio         5.2133     0.6401   8.145 2.11e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 54.73 on 33 degrees of freedom
Multiple R-squared:  0.8666,    Adjusted R-squared:  0.8585
F-statistic: 107.2 on 2 and 33 DF,  p-value: 3.689e-15
```

This model seems to be progressing! All the coefficients are significant, and the model as a whole is still significant. In fact, our adjusted R^2 has also improved slightly.

We will talk more about how to choose the best variables to build your regression models.

Also, don't forget that with every statistical procedure there are assumptions that need to be met. We can't correctly use either of these models yet because we haven't checked those assumptions. We'll will also discuss about how to do that in subsequent materials.