Cadreon R Test

Anna Andrianatou

27/04/2018

Install and loading of dependencies

```
list.of.packages <- c("readxl", "tidyverse", "GGally")
new.packages <- list.of.packages[!(list.of.packages %in% installed.packages()[,"Package"])]
if(length(new.packages)) install.packages(new.packages)

library(readxl)
library(tidyverse)
library(GGally)
```

Context

This is a R test given by Cadreon. It explores the sales of a toy brand across two years with respect to a monthly trend, digital and TV advertising investment and seasonal factors such as Christmas.

Question 1

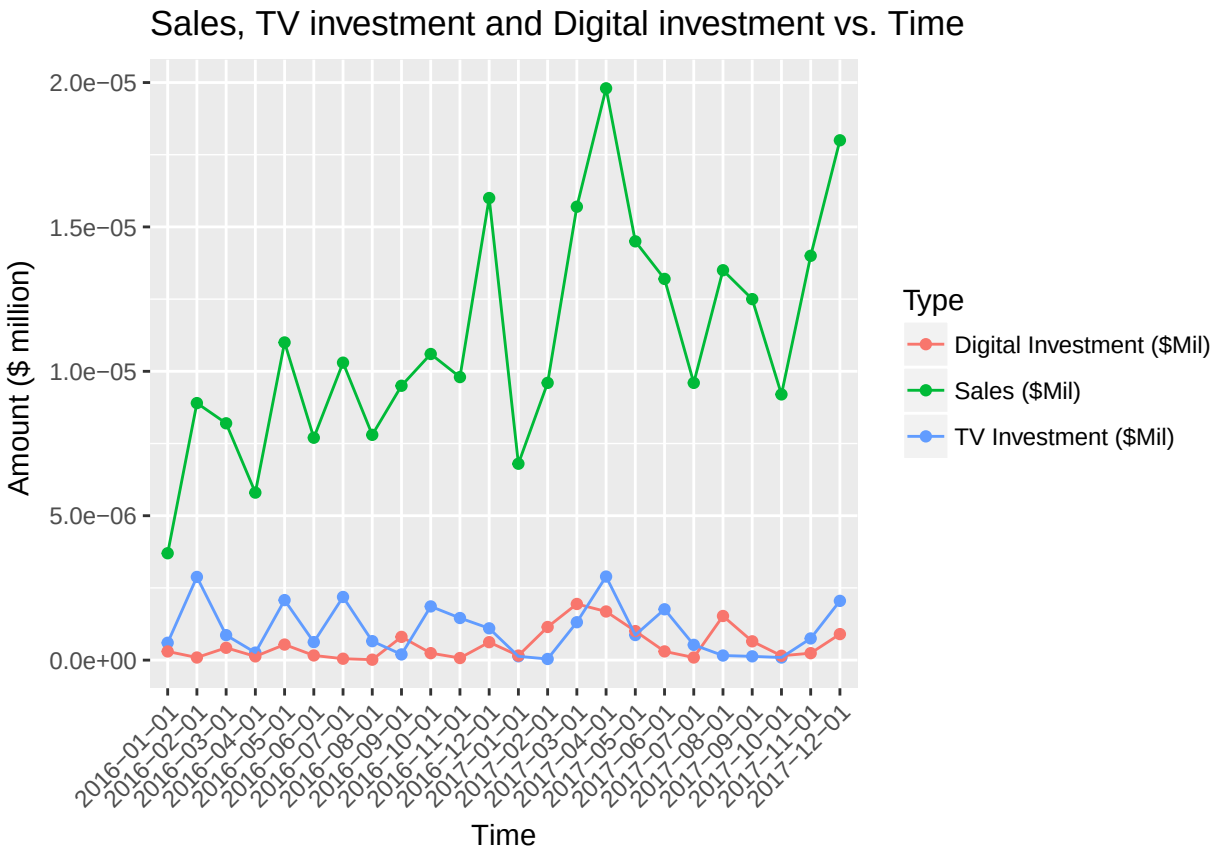
The following loads in the data and prepares it for plotting.

```
salesData = read_excel("../data/toy_sales_data.xlsx", sheet=1)
attach(salesData)

# Clean up the labels and scale the values for nice plotting
salesData <- salesData %>%
  mutate(digital_spend = digital_spend/10^6) %>%
  mutate(sales = sales/10^6) %>%
  mutate(tv_spend = tv_spend/10^6) %>%
  rename(`Digital Investment ($Mil)` = digital_spend) %>%
  rename(`Sales ($Mil)` = sales) %>%
  rename(`TV Investment ($Mil)` = tv_spend)
```

Question 2

A plot of sales, TV and digital investment is below. There appears to be a positive linear time trend in sales. There also appears to be some positive correlation between sales and both investment type.

```
tidySales = gather(salesData,  "Type", "Amount", 2:4)

ggplot(data= tidySales, aes(x=factor(month), y=Amount/10^6, colour=Type, group=Type))+
  geom_point()+
  geom_line()+
  ggtitle("Sales, TV investment and Digital investment vs. Time")+
  xlab("Time")+
  ylab("Amount ($ million) ") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



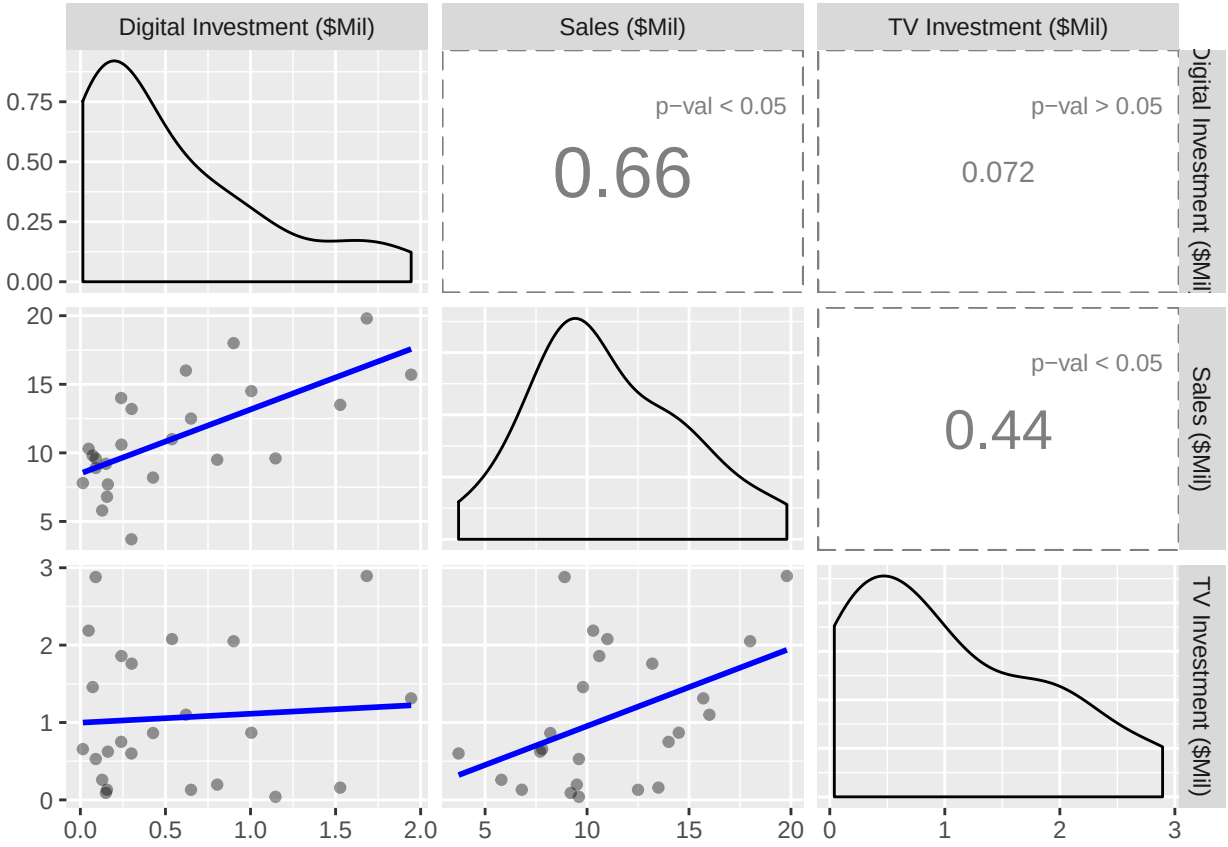
Question 3

A correlation matrix between sales, TV investment and digital investment is given below.

```
features <- c("Digital Investment ($Mil)", "Sales ($Mil)", "TV Investment ($Mil)")

lowerTriangleFunction <- function(data, mapping, ...){
  p <- ggplot(data = data, mapping = mapping) +
    geom_point(alpha = 0.4) +
    geom_smooth(method=lm, fill="blue", color="blue", se = FALSE, ...)
  p
}

currentTriangleFunction <- function(data, mapping, color = I("grey50"),
                                    sizeRange = c(3, 5), ...) {

  # get the x and y data to use the other code
  x <- eval(mapping$x, data)
  y <- eval(mapping$y, data)

  ct <- cor.test(x,y)
  sig <- symnum(
    ct$p.value,
    corr = FALSE,
    na = FALSE,
    cutpoints = c(0, 0.05, 0.1, 1),
```

```
    symbols = c("p-val < 0.05", ".", "p-val > 0.05")
  )

  r <- unname(ct$estimate)
  rt <- format(r, digits=2)[1]

  # since we can't print it to get the strsize, just use the max size range
  cex <- max(sizeRange)

  # helper function to calculate a useable size
  percent_of_range <- function(percent, range) {
    percent * diff(range) + min(range, na.rm = TRUE)
  }

  # plot the cor value
  ggally_text(
    label = as.character(rt),
    mapping = aes(),
    xP = 0.5, yP = 0.5,
    size = I(percent_of_range(cex * abs(r), sizeRange)),
    color = color,
    ...
  ) +
    # add the sig stars
    geom_text(
      aes_string(
        x = 0.8,
        y = 0.8
      ),
      label = sig,
      size = 3,
      color = color,
      ...
    ) +
    # remove all the background stuff and wrap it with a dashed line
    theme_classic() +
    theme(
      panel.background = element_rect(
        color = color,
        linetype = "longdash"
      ),
      axis.line = element_blank(),
      axis.ticks = element_blank(),
      axis.text.y = element_blank(),
      axis.text.x = element_blank()
    )
}

ggpairs(salesData %>%
          select(features),
        columns = features,
        lower = list(continuous = lowerTriangleFunction),
        upper = list(continuous = currentTriangleFunction)
```

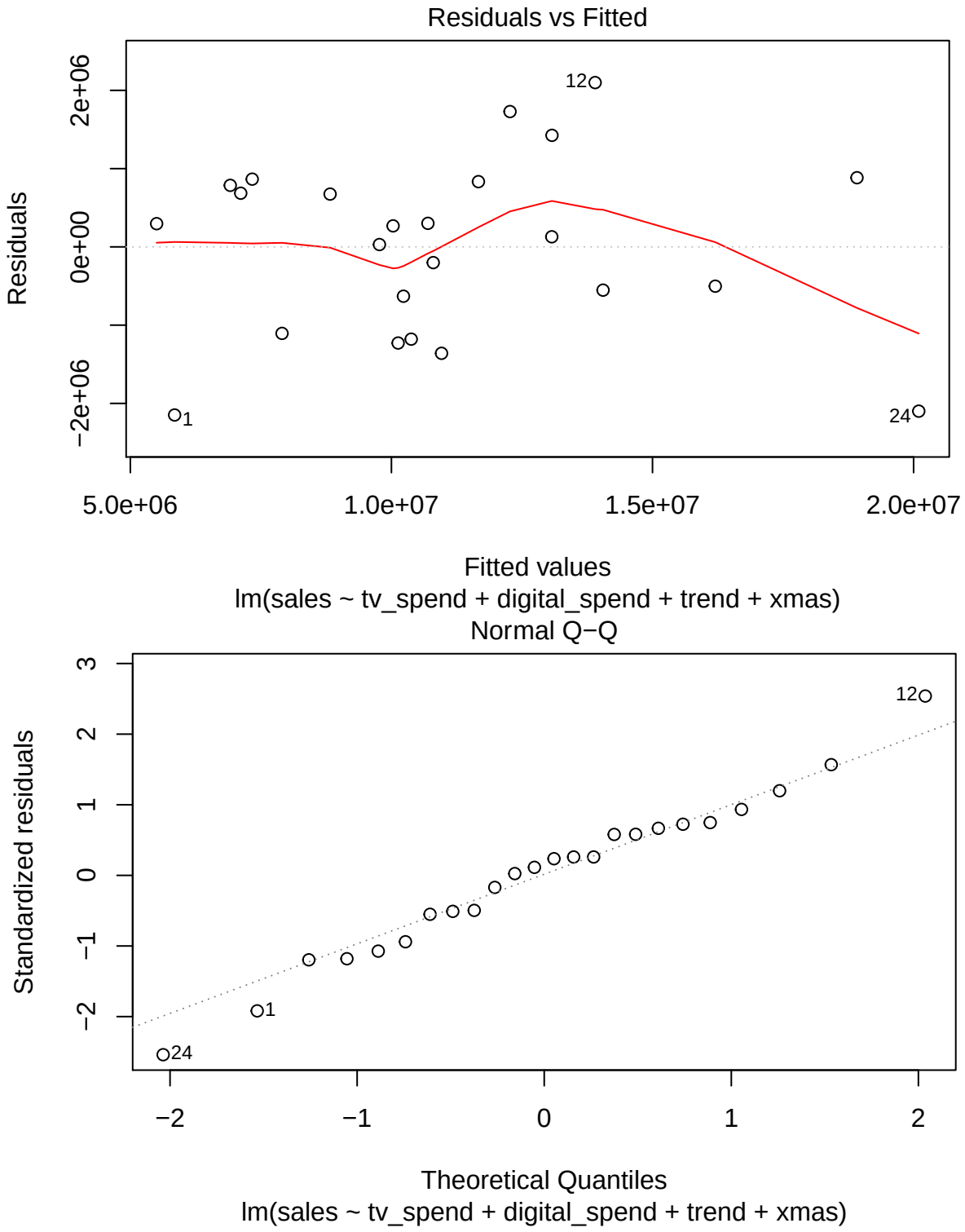
```
)
```



Sales is positively correlated with both types of investment. TV spending and digital spending do have some positive correlation, however is fairly small and not statistically significant at the 0.05 threshold. This is suggests that their relationship is unlikely to cause a multicolinearity problem.

Question 4 a,b

The month variable in R converts to an integer and represents time in seconds (the spacing between each month is different due to the different number of days in various months). The trend variable is an indicator for each month. We are only really interested in the monthly trend and so the "trend" variable is included and "month" omitted.




```
##
## Call:
## lm(formula = sales ~ tv_spend + digital_spend + trend + xmas)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -2148071  -749281   198668   798127  2099176
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.451e+06  6.706e+05   5.146 5.74e-05 ***
## tv_spend      2.026e+00  3.071e-01   6.599 2.58e-06 ***
## digital_spend 2.983e+00  5.040e-01   5.919 1.07e-05 ***
## trend         2.863e+05  4.156e+04   6.888 1.44e-06 ***
## xmas          2.935e+06  9.730e+05   3.016  0.00711 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1245000 on 19 degrees of freedom
## Multiple R-squared:  0.9161, Adjusted R-squared:  0.8984
## F-statistic: 51.84 on 4 and 19 DF,  p-value: 5.817e-10
```

The adjusted R squared (adjusted for the number of variables in the model) is 0.8984. Therefore, approximately 90% of the variation in the sales data is explained by the model.
Assume a level of significance of 5%:
TV spend regressor has an associated p-value of 2.58e-06 (<0.05) and therefore there is statisitcally significant in the model (coefficient $\neq 0$).

Digital spend regressor has an associated p-value of 1.07e-05 and therefore there is statisitcally significant in the model (coefficient $\neq 0$).

Trend has an associated p-value of 1.44e-06 and therefore there is statisitcally significant in the model (coefficient $\neq 0$). There is some growth over time.

xmas has an associated p-value of 0.00711 and therefore there is statisitcally significant in the model (coefficient $\neq 0$). X-mas effects sales positively.

Question 5

From this regression, for every dollar in TV_spend, sales increase by $2.026.
To find percentage change, fit a log-linear regression and interpret the coefficient of TV_spend

```
logSalesRegression = lm(log(sales) ~ tv_spend + digital_spend + trend + xmas)
summary(logSalesRegression)
```

```
##
## Call:
## lm(formula = log(sales) ~ tv_spend + digital_spend + trend +
##     xmas)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50464 -0.05648  0.01681  0.10188  0.25602
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.541e+01  9.255e-02 166.503  < 2e-16 ***
## tv_spend      1.957e-07  4.238e-08   4.618 0.000188 ***
## digital_spend 2.310e-07  6.955e-08   3.321 0.003590 **
## trend         3.158e-02  5.737e-03   5.506  2.6e-05 ***
## xmas          1.843e-01  1.343e-01   1.372 0.185995
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1718 on 19 degrees of freedom
## Multiple R-squared:  0.8308, Adjusted R-squared:  0.7951
## F-statistic: 23.32 on 4 and 19 DF,  p-value: 4.167e-07
```

From this log-linear regression, for every dollar in TV_spend, sales increase by 0.00001957%.

Question 6

$ROI = (gains - investmentcost)/investmentcost$ ($2.026 for every $1 in investment)

```
ROI = (2.026-1)/1*100    #102.6%
```

Question 7

Subsitute the new data into the fitted model. The fitted model is:
$sales = 3.451 \times 10^6 + 2.026 TV_{spend} + 2.983 Digital_{spend} + 2.863 \times 10^5 t + 2.935 \times 10^6 \mathbb{I}(Xmas)$.
Plug in the 3 years of planned investment, with time trend $(25, 26, 27)$ and Xmas is $(0, 0, 0)$ since none of the months are December. The predictions are made using predict.lm.

```
plannedData = read_excel("../data/toy_sales_data.xlsx", sheet=2)
plannedData$trend = c(25,26,27)  #continue the trend variable for the following months
plannedData$xmas = c(0,0,0)      #none of the prediction dates are xmas
predict.lm(salesRegression, plannedData)   #predicted sales
```

```
##        1        2        3
## 11958795 13267129 15109712
```

Question 8

The purpose of the model is to observe the return to investment of TV and digital ads and to predict future sales. Therefore, any additional data that could explain sales would be beneficial to the model. This could include any other important events other than xmas that could increase the sales of toys, prices or sales of competing toy brands or subsitution items(such as video games), the number of stores selling the toy brand in each month.

It is important to capture the impact of effects other than the advertising campaign to ensure the impact of the ads are not over estimated.