Lab 9

2023-07-20

Introduction

Download the data set media_spend.csv and save it to whatever directory you are using for this course.

Today's dataset contains information about a fictitious company that's trying to determine how much money to spend on various types of advertising for the coming year. They have historical data showing sales (in millions of dollars) and the amount they spent on TV, Radio, and Newspaper advertising that year (in thousands of dollars). Our goal today is to determine which types of advertising effects sales the most.

Question 1

```
library(dplyr)
#(a) Load the media spend.csv dataset into R and save it to spend.
spend <- read_csv('media_spend.csv')
```

```
## Rows: 200 Columns: 4
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## dbl (4): TV, Radio, Newspaper, Sales
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#(b) We are trying to see how the advertising effects sales, therefore, sales is our
response variable and the other columns are our regressors. List the regressors (inde
pendent variables).
regressors <- names(spend)[names(spend) != 'Sales']
response <- names(spend)[names(spend) == 'Sales']

regressors
```

```
## [1] "TV"        "Radio"     "Newspaper"
```

```
response
```

```
## [1] "Sales"
```

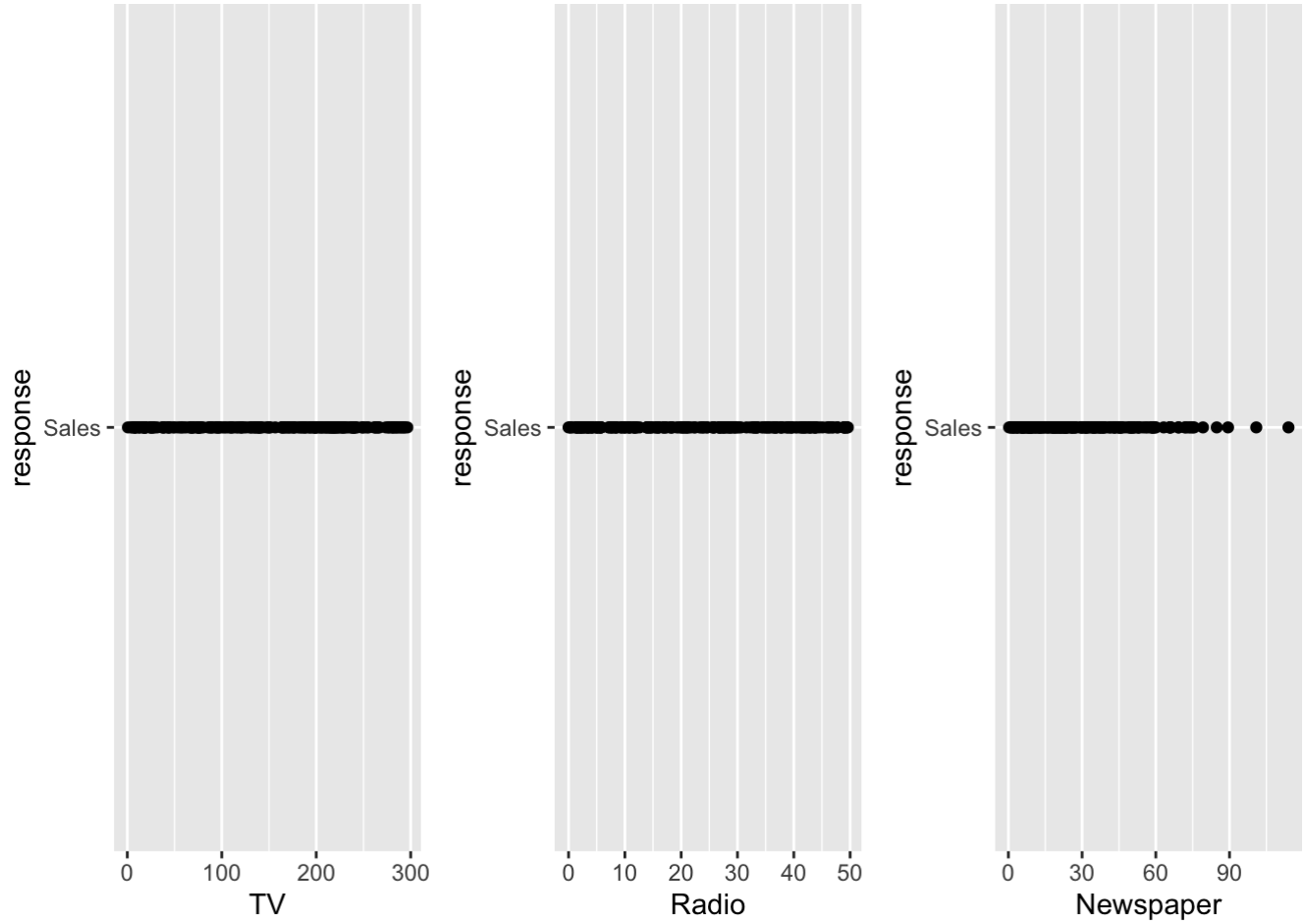
```
#(c) Plot the response variable (as the y-axis) against each of the regressor variabl
es (one plot for each regressor) using ggplot2. Display the three plots side by side
using the "gridExtra" package.

#install.packages("gridExtra")              # install package if needed
library("gridExtra")                        # Load gridExtra package
```

```
##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
createPlot <- function(regressor) {
  ggplot(spend, aes(x = .data[[regressor]], y = response)) +
  geom_point()
}

plots <- lapply(regressors, createPlot)
grid.arrange(grobs = plots, ncol = 3)
```




```
#(d) Looking only at the plots, which type of advertising do you think will have the
largest effect on sales?
# The media with the largest effect on sales is by far TV because it has the highest
number coming up at 300.
```

The media with the largest effect on sales is by far TV because it has the highest number coming up at 300.

Question 2

```
#(a) Perform a linear regression for each form of advertising vs the response variabl
e, sales. Print out the summary for each of these regressions and take note of the p-
value for the t-test on the significance of the coefficient for each.

target <- spend$Sales



tv_summary <- summary(lm(formula = target ~ spend$TV, data = spend))
radio_summary <- summary(lm(formula = target ~ spend$Radio, data = spend))
newspaper_summary <- summary(lm(formula = target ~ spend$Newspaper, data = spend))

tv_summary
```

```
##
## Call:
## lm(formula = target ~ spend$TV, data = spend)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4438 -1.4857  0.0218  1.5042  5.6932
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.974821   0.322553   21.62   <2e-16 ***
## spend$TV    0.055465   0.001896   29.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.296 on 198 degrees of freedom
## Multiple R-squared:  0.8122, Adjusted R-squared:  0.8112
## F-statistic: 856.2 on 1 and 198 DF,  p-value: < 2.2e-16
```

```
radio_summary
```

```
##
## Call:
## lm(formula = target ~ spend$Radio, data = spend)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.5632  -3.5293   0.6714   4.2504   8.6796
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.2357     0.6535  18.724  < 2e-16 ***
## spend$Radio   0.1244     0.0237   5.251 3.88e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.963 on 198 degrees of freedom
## Multiple R-squared:  0.1222, Adjusted R-squared:  0.1178
## F-statistic: 27.57 on 1 and 198 DF,  p-value: 3.883e-07
```

```
newspaper_summary
```

```
##
## Call:
## lm(formula = target ~ spend$Newspaper, data = spend)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.6930  -3.8807   0.6591   3.9083  11.4385
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     13.95955    0.63829  21.870   <2e-16 ***
## spend$Newspaper  0.03832    0.01703   2.251   0.0255 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.231 on 198 degrees of freedom
## Multiple R-squared:  0.02495,    Adjusted R-squared:  0.02003
## F-statistic: 5.067 on 1 and 198 DF,  p-value: 0.02549
```

```
#(b) Which of the regressors is the most significant in terms of the p-value?

# The regressor with the most significance in terms of the p-value
# is TV meaning that it has a strong response with the 'Sales' response variable.

#(c) Of all the significant regressors, which one has the biggest impact on sales on
its own?
# On its own, the regressor with the biggest impact is Radio because it has the bigge
st unit increase in sales (0.1244)

#(d) Create a vector of the p-values and name each element with the corresponding typ
e of advertising.
lmp <- function (summaryObject) {
    if (class(summaryObject) != "summary.lm") stop("Not an object of class 'summary.l
m' ")
    f <- summaryObject$fstatistic
    p <- pf(f[1],f[2],f[3],lower.tail=F)
    attributes(p) <- NULL
    return(p)
}

summaries <- c(tv_summary, radio_summary, newspaper_summary)

pValues <- c(lmp(tv_summary), lmp(radio_summary), lmp(newspaper_summary))

headers <- c('TV', 'Radio', 'NewsPaper')

names(pValues) <- headers




#(e) Create another vector of the slope coefficients and name each element with the c
orresponding type of advertising.
slope_coefficients <- c(tv_summary$coefficients[,"Estimate" ], radio_summary$coeffici
ents[,"Estimate" ], newspaper_summary$coefficients[,"Estimate" ])

names(slope_coefficients) <- headers



#(f) Display both vectors.
pValues
```

```
##           TV        Radio    NewsPaper
## 7.927912e-74 3.882892e-07 2.548744e-02
```

```
slope_coefficients
```

```
##          TV       Radio   NewsPaper        <NA>        <NA>        <NA>
##  6.97482149  0.05546477 12.23572197  0.12443166 13.95954865  0.03832400
```

b. Which of the regressors is the most significant in terms of the p-value?

The regressor with the most significance in terms of the p-value is TV meaning that it has a strong response with the ‘Sales’ response variable.

End of Lab 9.