Lab 10: Multiple Regression

2023-07-27

Introduction

We will continue our analysis of the fictitious company that's trying to determine how much money to spend on various types of advertising for the coming year. If you don't have your data file "media spend.csv" from the last lab in your working directory, download it from Brightspace.

Question 1

```
#(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.
```

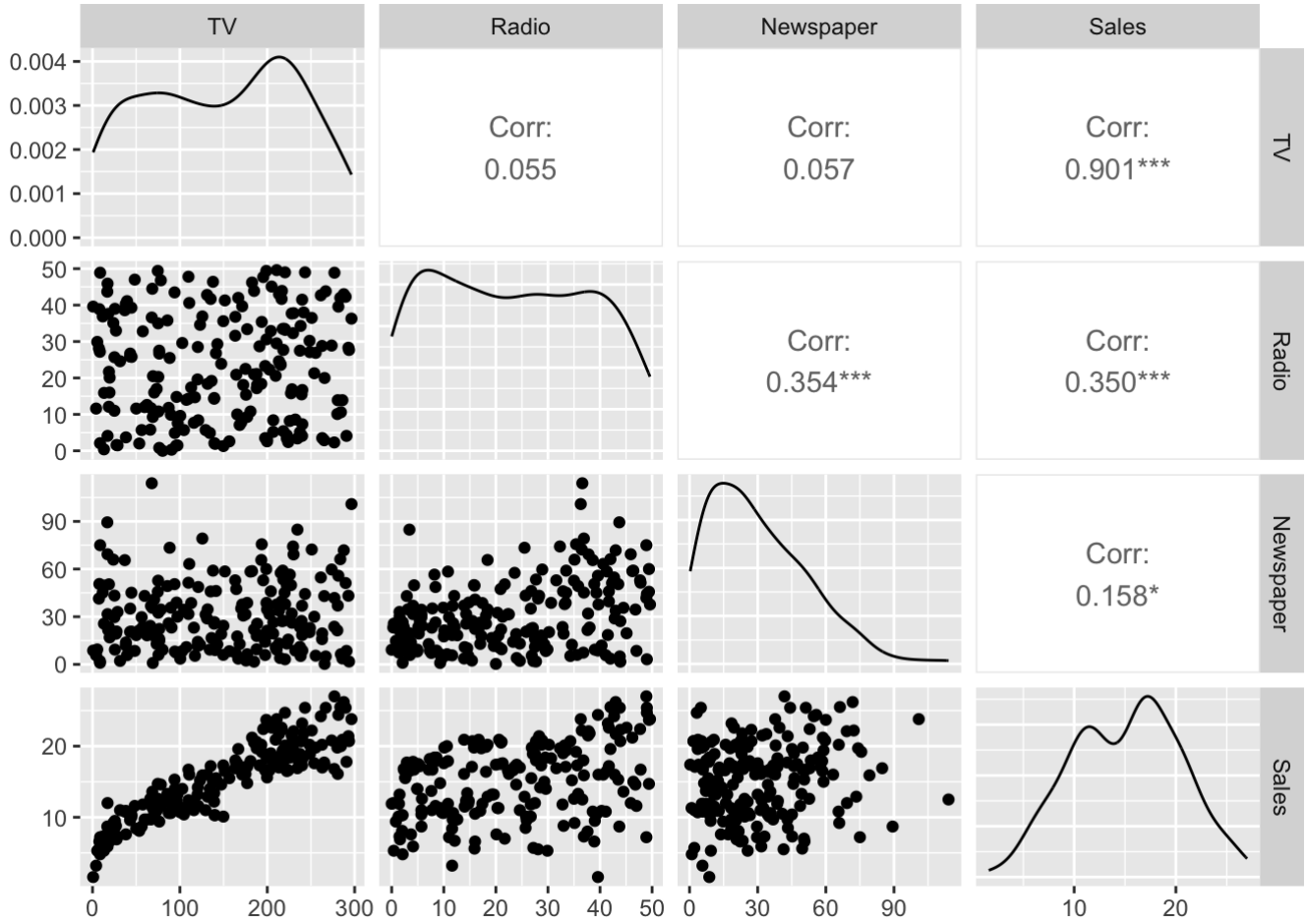
```
#(b) Install and load GGally package

#install.packages("GGally")    # only run once
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
#(c)  Produce a correlation matrix of all variables using the command ggpairs(paramet
ers).
#     Since we want to show all variables in the spend data frame, we can input spend
as the
#     parameter.
args(ggpairs)
```

```
## function (data, mapping = NULL, columns = 1:ncol(data), title = NULL,
##     upper = list(continuous = "cor", combo = "box_no_facet",
##         discrete = "count", na = "na"), lower = list(continuous = "points",
##         combo = "facethist", discrete = "facetbar", na = "na"),
##     diag = list(continuous = "densityDiag", discrete = "barDiag",
##         na = "naDiag"), params = NULL, ..., xlab = NULL, ylab = NULL,
##     axisLabels = c("show", "internal", "none"), columnLabels = colnames(data[colum
ns]),
##     labeller = "label_value", switch = NULL, showStrips = NULL,
##     legend = NULL, cardinality_threshold = 15, progress = NULL,
##     proportions = NULL, legends = stop("deprecated"))
## NULL
```

```
ggpairs(spend)
```




```
#(d)  Which pair(s) of independent variables showed significant correlation (+ or -)?

# Answer:
```

TV advertising spending has a strong positive relationship with Sales, while Radio and Newspaper advertising spending have weaker positive relationships with Sales. However, the correlations between advertising spending in different channels (TV, Radio, Newspaper) are generally weak, indicating that they are less related to each other.

Question 2

```
#(a) Perform a multiple regression using all three types of advertising vs the
#    response variable, Sales. Save the results and name it full_model. Print out
#    the summary of the regression.

full_model <- lm(Sales ~ TV + Radio + Newspaper, data = spend)
summary(full_model)
```

```
##
## Call:
## lm(formula = Sales ~ TV + Radio + Newspaper, data = spend)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3034 -0.8244 -0.0008  0.8976  3.7473
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.6251241  0.3075012  15.041   <2e-16 ***
## TV          0.0544458  0.0013752  39.592   <2e-16 ***
## Radio       0.1070012  0.0084896  12.604   <2e-16 ***
## Newspaper   0.0003357  0.0057881   0.058    0.954
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.662 on 196 degrees of freedom
## Multiple R-squared:  0.9026, Adjusted R-squared:  0.9011
## F-statistic: 605.4 on 3 and 196 DF,  p-value: < 2.2e-16
```

```
#(b) Compare and comment on the slope coefficients of the explanatory variables and t
heir
#    corresponding p-values with those obtained from your previous lab when you perfo
rmed
#    linear regressions on individual explanatory variables.


# Here are values from the previous from last lab.

slope_vec <- c(0.055465, 0.1244, 0.03832)
names(slope_vec) <- c("TV", "Radio", "Newspaper")

p_vec <- c(2.2e-16, 3.88e-07, 0.02549)
names(p_vec) <- c("TV", "Radio", "Newspaper")

########################
slope_full <- coef(full_model)
p_values_full <- summary(full_model)$coefficients[, 4]

#slope coefficients
for(i in names(slope_vec)) {
 cat(i, " | individual: ", slope_vec[i], " | full: ", slope_full[i], "\n")
}
```

```
## TV  | individual:  0.055465  | full:  0.05444578
## Radio  | individual:  0.1244  | full:  0.1070012
## Newspaper  | individual:  0.03832  | full:  0.0003356579
```

```
#p-values
for (var in names(p_vec)) {
  cat(var, " | individual: ", p_vec[var], " | full: ", p_values_full[var], "\n")
}
```

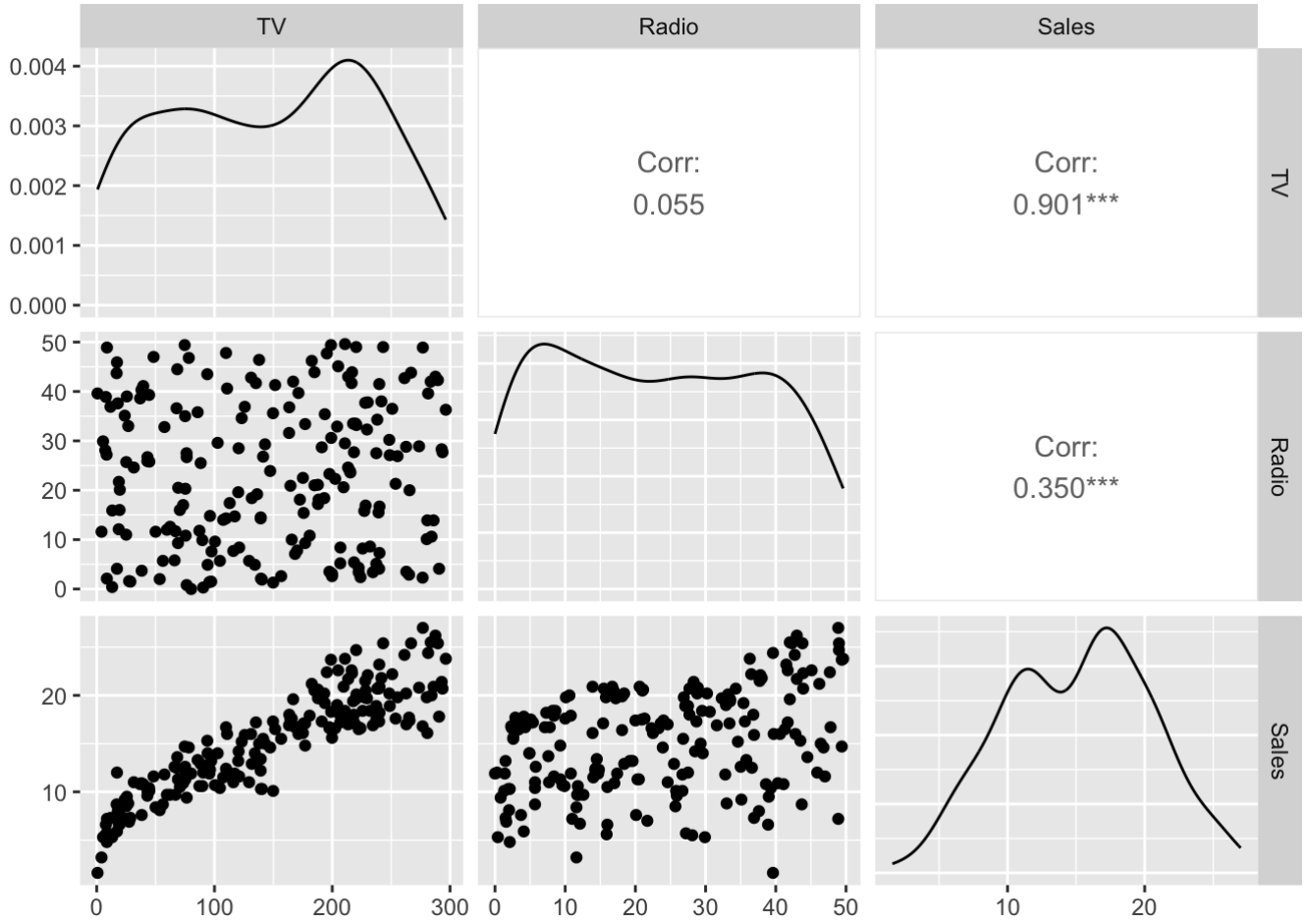
```
## TV  | individual:  2.2e-16  | full:  1.892945e-95
## Radio  | individual:  3.88e-07  | full:  4.602097e-27
## Newspaper  | individual:  0.02549  | full:  0.9538145
```

```
# TV advertising is highly significant and has consistent positive effect on Sales.
# Radio advertising is highly significant and has a slightly reduced positive effect
on Sales.
# Newspaper advertising, appeared significant in the individual regression, becomes i
nsignificant when considering other advertising types.


########################

# Note: Once you introduce a multiple regression model, the number in the
# individual regression analysis should be only used for comparison purpose.

# (c) Which of the regressor(s) is(are) not statistically significant in the full mod
el?
#     Explain how you know they are not significant.

# Answer:
```

based on the p-values in the full model regression summary, “Newspaper” is not statistically significant in predicting Sales when considering the effects of TV and Radio advertising.

Question 3

```
# Optional: You can "clean up" your correlation by skipping variables you
# don't want to use.

# Uncomment it if you want to see the output.

ggpairs(subset(spend, select = c("TV", "Radio", "Sales")))
```




```
# The subset subset command includes the variables you specify.

# (a) Run a multiple regression with only the regressors that are significant
#     in full model and name it new model. Print out the summary.

new_model <- lm(Sales ~ TV + Radio, data = spend)
summary(new_model)
```

```
##
## Call:
## lm(formula = Sales ~ TV + Radio, data = spend)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3131 -0.8269  0.0095  0.9022  3.7484
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.630879   0.290308   15.95   <2e-16 ***
## TV          0.054449   0.001371   39.73   <2e-16 ***
## Radio       0.107175   0.007926   13.52   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.657 on 197 degrees of freedom
## Multiple R-squared:  0.9026, Adjusted R-squared:  0.9016
## F-statistic: 912.7 on 2 and 197 DF,  p-value: < 2.2e-16
```

```
#(b) Are there any insignificant regressors in this model?

# Answer:

# No, all the regressors have low p-values indicating a significantly different coeff
icient
# from 0.


#(c) Now you have two regression models: full_model and new_model. Which one will you
#    select as the final (better) regression model. Justify your choice.

# Answer:
# Overall, the new_model has a better balance by including only the significant predi
ctors "TV" and "Radio," making it a better choice for explaining the variability in S
ales while avoiding potential issues associated with non-significant predictors.


#(d) Write out the equation for new_model.

# Answer:

# Sales = 4.630879 + 0.054449 * TV + 0.107175 * Radio

# Recall The unit for Sales is millions of dollars and the units for TV, Radio,
# and Newspaper are in thousands of dollars.

#(e) Use new_model to predict sales if the company spends $200,000 in TV ad and
#    $30,000 in radio ad. Is this prediction a good one for this model? Why or
#    why not?


# The predicted sales is

predicted_sales <- predict(new_model, newdata = data.frame(TV = 200000, Radio = 3000
0))
predicted_sales
```

```
##        1
## 14109.66
```

```
# it appears to be a reasonable prediction, given the adjusted R-squared value of 0.9
016 in the new_model.

#(f) Use new_model to predict sales if the company spends $500,000 in TV ad and
#    $80,000 in radio ad. Is this prediction a good one for this model? Why or
#    why not?

predicted_sales2 <- predict(new_model, newdata = data.frame(TV = 500000, Radio = 8000
0))
# The predicted sales is
predicted_sales2
```

```
##        1
## 35803.08
```

```
# based value of $35,803.08, it appears to be a relatively high Sales prediction give
n the spending amounts of $500,000 on TV ads and $80,000 on Radio ads.
```

Knit this file.

Open the html file in your browser.

Print that file as a pdf file.

Submit the pdf file to Brightspace.

End of Lab 10.