problemset-1b.R

Aaryan Agarwal

2023-09-12

```
# This assignment is done in a group of three
# Prajwal Kaushal, Sumukha Sharma, Aaryan Agarwal

# Part a
#Question 1
data=read.csv("D:/Programs/Personal/MSDS/Machine Learning with stats data/Advertising.csv")
data[1:5,]
```

```
##   X    TV Radio Newspaper Sales
## 1 1 230.1  37.8      69.2  22.1
## 2 2  44.5  39.3      45.1  10.4
## 3 3  17.2  45.9      69.3   9.3
## 4 4 151.5  41.3      58.5  18.5
## 5 5 180.8  10.8      58.4  12.9
```

```
#removing the first row as it is just the index
data=data[,2:5]
data[1:5,]
```

```
##      TV Radio Newspaper Sales
## 1 230.1  37.8      69.2  22.1
## 2  44.5  39.3      45.1  10.4
## 3  17.2  45.9      69.3   9.3
## 4 151.5  41.3      58.5  18.5
## 5 180.8  10.8      58.4  12.9
```

```
#summary of data
summary(data)
```

```
##        TV             Radio          Newspaper          Sales      
##  Min.   :  0.70   Min.   : 0.000   Min.   :  0.30   Min.   : 1.60  
##  1st Qu.: 74.38   1st Qu.: 9.975   1st Qu.: 12.75   1st Qu.:10.38  
##  Median :149.75   Median :22.900   Median : 25.75   Median :12.90  
##  Mean   :147.04   Mean   :23.264   Mean   : 30.55   Mean   :14.02  
##  3rd Qu.:218.82   3rd Qu.:36.525   3rd Qu.: 45.10   3rd Qu.:17.40  
##  Max.   :296.40   Max.   :49.600   Max.   :114.00   Max.   :27.00  
```

```
#plotting the data
plot(data)
```

```
#Question 2
#Linear Regression between Sales and TV
model1=lm(Sales~TV, data = data)
summary(model1)
```

```
##
## Call:
## lm(formula = Sales ~ TV, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.032594   0.457843   15.36   <2e-16 ***
## TV          0.047537   0.002691   17.67   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

```
#R sq 0.609 is a good fir
#Linear Regression between Sales and Radio
model2=lm(Sales~Radio, data = data)
summary(model2)
```

```
##
## Call:
## lm(formula = Sales ~ Radio, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7305  -2.1324   0.7707   2.7775   8.1810
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.31164    0.56290  16.542   <2e-16 ***
## Radio        0.20250    0.02041   9.921   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.275 on 198 degrees of freedom
## Multiple R-squared:  0.332,  Adjusted R-squared:  0.3287
## F-statistic: 98.42 on 1 and 198 DF,  p-value: < 2.2e-16
```

```
#R sq 0.328 is a avg fit
#Linear Regression between Sales and Newspaper
model3=lm(Sales~Newspaper, data = data)
summary(model3)
```

```
## 
## Call:
## lm(formula = Sales ~ Newspaper, data = data)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -11.2272  -3.3873  -0.8392   3.5059  12.7751 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 12.35141    0.62142   19.88  < 2e-16 ***
## Newspaper    0.05469    0.01658    3.30  0.00115 ** 
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 5.092 on 198 degrees of freedom
## Multiple R-squared:  0.05212,    Adjusted R-squared:  0.04733 
## F-statistic: 10.89 on 1 and 198 DF,  p-value: 0.001148
```

```
#R sq 0.047 is a bad fit



#Question 3
#Multiple Linear Regression
model4=lm(Sales~TV+Newspaper+Radio,data=data)
summary(model4)
```

```
## 
## Call:
## lm(formula = Sales ~ TV + Newspaper + Radio, data = data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -8.8277 -0.8908  0.2418  1.1893  2.8292 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  2.938889   0.311908   9.422   <2e-16 ***
## TV           0.045765   0.001395  32.809   <2e-16 ***
## Newspaper   -0.001037   0.005871  -0.177     0.86    
## Radio        0.188530   0.008611  21.893   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956 
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

```
#R sq 0.895 is a good fit
library(rgl)
```

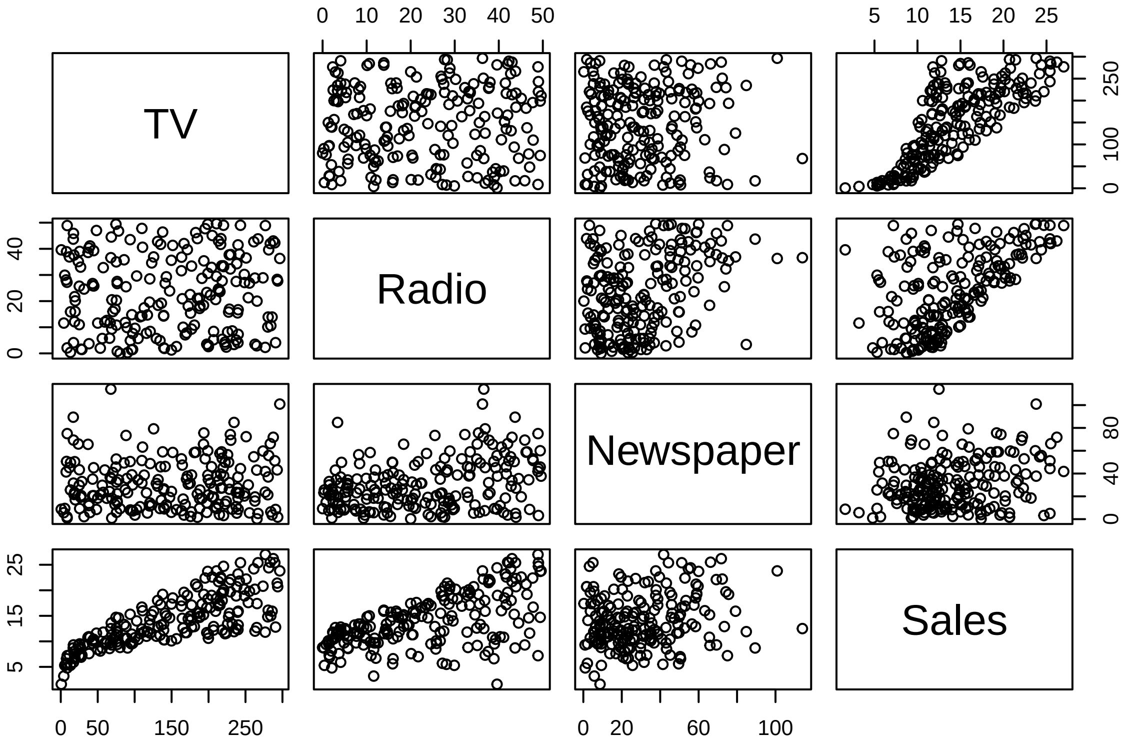
```
## Warning: package ’rgl’ was built under R version 4.2.3
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.2.3

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.2.2
```

```
scatter3d(Sales~TV+Radio,data=data)
```

```
## Loading required namespace: mgcv

## Loading required namespace: MASS
```



```
#Question 4
# Multiple Regression with Interaction Term between TV and Radio
model5=lm(Sales ~ TV * Radio, data=data)
summary(model5)
```

```
##
## Call:
## lm(formula = Sales ~ TV * Radio, data = data)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3366 -0.4028  0.1831  0.5948  1.5246
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
## TV          1.910e-02  1.504e-03  12.699   <2e-16 ***
## Radio       2.886e-02  8.905e-03   3.241   0.0014 **
## TV:Radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.9435 on 196 degrees of freedom
## Multiple R-squared:  0.9678, Adjusted R-squared:  0.9673
## F-statistic:  1963 on 3 and 196 DF,  p-value: < 2.2e-16
```

```
#R sq 0.967 is a very good fit
#Better fit than previous models

#Other Interaction terms
model6=lm(Sales ~ TV * Newspaper, data=data)
summary(model6)
```

```
##
## Call:
## lm(formula = Sales ~ TV * Newspaper, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1860 -1.5521 -0.0648  1.8062  8.7276
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.4042175  0.7333818   8.732  1.1e-15 ***
## TV           0.0426585  0.0043105   9.896  < 2e-16 ***
## Newspaper    0.0241103  0.0192716   1.251    0.212
## TV:Newspaper 0.0001324  0.0001079   1.228    0.221
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.117 on 196 degrees of freedom
## Multiple R-squared:  0.6485, Adjusted R-squared:  0.6432
## F-statistic: 120.6 on 3 and 196 DF,  p-value: < 2.2e-16
```

```
#R sq 0.643 good fit

model7=lm(Sales ~ Radio * Newspaper, data=data)
summary(model7)
```

```
##
## Call:
```

```
## lm(formula = Sales ~ Radio * Newspaper, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.6981  -2.1955   0.7567   2.7191   8.2228
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      8.7904734  1.0224848   8.597 2.58e-15 ***
## Radio            0.2145684  0.0382985   5.603 7.08e-08 ***
## Newspaper        0.0220611  0.0345866   0.638    0.524
## Radio:Newspaper -0.0005259  0.0010642  -0.494    0.622
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.292 on 196 degrees of freedom
## Multiple R-squared:  0.3335, Adjusted R-squared:  0.3233
## F-statistic:  32.7 on 3 and 196 DF,  p-value: < 2.2e-16
```

```
#R sq 0.33 avg fit

model8=lm(Sales ~ TV * Radio * Newspaper, data=data)
summary(model8)
```

```
##
## Call:
## lm(formula = Sales ~ TV * Radio * Newspaper, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8955 -0.3883  0.1938  0.5865  1.5240
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         6.556e+00  4.655e-01  14.083  < 2e-16 ***
## TV                  1.971e-02  2.719e-03   7.250 9.95e-12 ***
## Radio               1.962e-02  1.639e-02   1.197    0.233
## Newspaper           1.311e-02  1.721e-02   0.761    0.447
## TV:Radio            1.162e-03  9.753e-05  11.909  < 2e-16 ***
## TV:Newspaper       -5.545e-05  9.326e-05  -0.595    0.553
## Radio:Newspaper     9.063e-06  4.831e-04   0.019    0.985
## TV:Radio:Newspaper -7.610e-07  2.700e-06  -0.282    0.778
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.9406 on 192 degrees of freedom
## Multiple R-squared:  0.9686, Adjusted R-squared:  0.9675
## F-statistic: 847.3 on 7 and 192 DF,  p-value: < 2.2e-16
```

```
#R sq 0.967 is a very good fit

#Question 5
#Linear Regression between Sales and (TV+Radio+TV:Radio)
```

```
model9=lm(Sales ~ TV + Radio + TV:Radio , data=data)
summary(model9)
```

```
##
## Call:
## lm(formula = Sales ~ TV + Radio + TV:Radio, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3366 -0.4028  0.1831  0.5948  1.5246
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
## TV          1.910e-02  1.504e-03  12.699   <2e-16 ***
## Radio       2.886e-02  8.905e-03   3.241   0.0014 **
## TV:Radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9435 on 196 degrees of freedom
## Multiple R-squared:  0.9678, Adjusted R-squared:  0.9673
## F-statistic:  1963 on 3 and 196 DF,  p-value: < 2.2e-16
```

```
beta1=coef(model9)["TV"]
beta1
```

```
##         TV
## 0.01910107
```

```
beta2=coef(model9)["Radio"]
beta2
```

```
##      Radio
## 0.02886034
```

```
beta3=coef(model9)["TV:Radio"]
beta3
```

```
##    TV:Radio
## 0.001086495
```

```
# optimal values of TV and Radio
newTV=(beta1-beta2+300*beta3)/(2*beta3)
newTV
```

```
##       TV
## 145.5088
```

```
newRadio=300-newTV
newRadio
```

```
##       TV
## 154.4912
```

```
newdata=data.frame(TV=newTV,Radio=newRadio)
newdata
```

```
##          TV    Radio
## TV 145.5088 154.4912
```

```
#optimal sales
newSales=predict(model9,newdata = newdata)
newSales
```

```
##       TV
## 38.41248
```

```
#confidence interval for the prediction
ci=predict(model9, newdata=newdata, interval="confidence")
ci
```

```
##         fit      lwr     upr
## TV 38.41248 37.23716 39.5878
```

```
# Part B

# What is the goal of Machine Learning?
# Machine learning specialists are often primarily concerned with developing
# high-performance computer systems that can provide useful predictions in the
# presence of challenging computational constraints.
# It offers a set of tools that can usefully summarize various sorts
# of nonlinear relationships in the data.
# The goal of machine learning is typically to achieve good out-of-sample predictions.
# In other words, the aim is to build models that perform well on new, unseen data,
# rather than just on the data they were trained on.

####################################################################################
# What does Varian mean by "good out-of-sample predictions"?
# "Good out-of-sample predictions" refers to the ability of a model to generalize
# well to new data that it hasn't seen before (data that was not used in training the model).
# A model that makes accurate predictions on new data has good out-of-sample performance

####################################################################################
# What is overfitting?
# Overfitting occurs when a model is too complex and fits the training data too closely,
# capturing even its noise. Such a model will perform poorly on new, unseen data
# because it has become too tailored to the training set.

####################################################################################
```

```
# What is model complexity?
# Model complexity refers to the number of parameters in a model or the intricacy of its structure.
# A more complex model might fit the training data very well but may not generalize well to new
# data,leading to overfitting. Varian suggests that if we have a numeric measure of model complexity,
# we can view it as a parameter that can be adjusted or "tuned" to achieve the best out-of-sample
# predictions.

####################################################################################
# What is the training data?
# Training data is used to estimate or train a model. In the process of building a model,
# data is typically split into training, validation, and testing sets.
# The model is trained on the training data, the best model structure or hyperparameters are
# chosen using the validation data, and the model's performance is evaluated on the testing data.
```