E392: Problem Set 4. Solutions

Linear Regression

Spring 2018

Due: February 26 2018

Please work on the following questions and hand in your solutions in groups of at most 2 students. You are asked to answer all questions, but we will only select 2 (sub)questions randomly to grade.

Suppose you are hired as a consultant to provide advice on how to improve sales of a particular product. The Advertising data set consists of the sales (in thousands of units) of that product in 200 different markets, along with advertising budgets (in thousands of dollars) for the product in each of those markets for three different media: TV, radio, and newspaper. Our goal is to develop an accurate model that can be used to predict sales on the basis of the three media budgets and obtain an optimal allocation of budgets based on the model. Your objective is to analyze this data set with R to answer your client's questions. The data set can be found in Canvas (Advertising.cvs).

Part 1: R questions

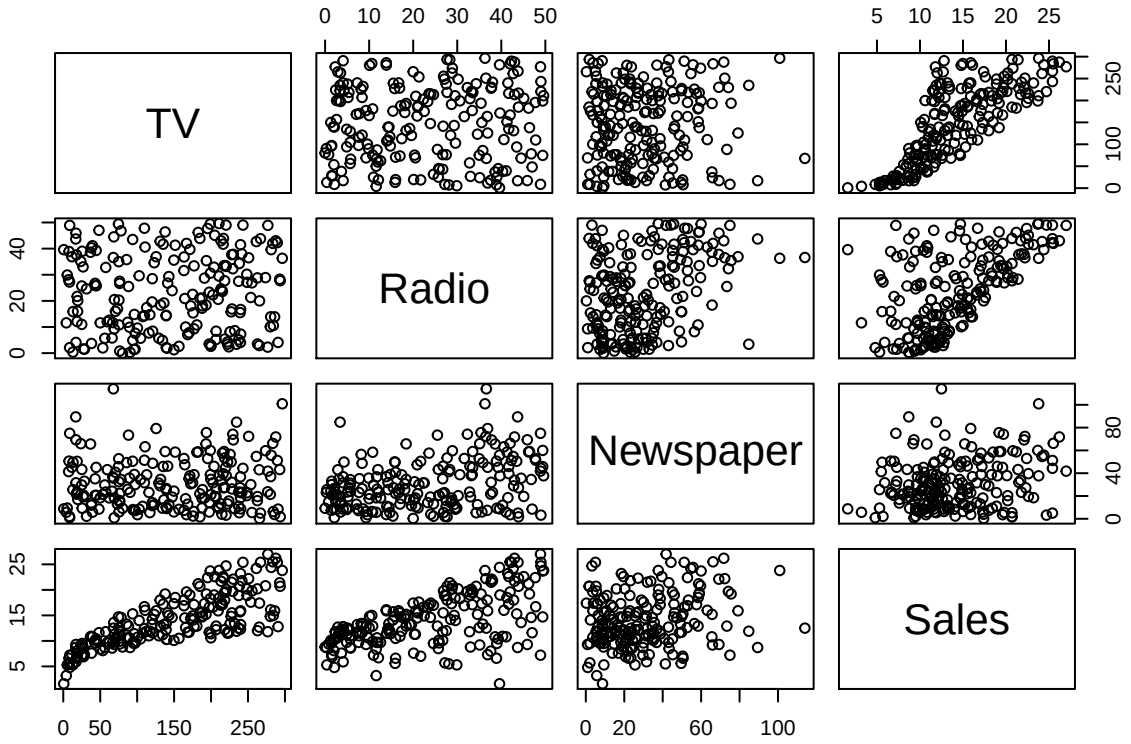
Question 1: Load and visualize the data

Load the data using the command `read.cvs`, summarize the data, and plot the data.

```
setwd("C:/Users/jescanci/Dropbox/teaching/2017-2018/e390-bigdata/ProblemSets")
Advertising=read.csv("Advertising.csv")
Advertising <- subset(Advertising, select = -c(1) )
summary(Advertising)
```

```
##        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
```

```
pairs(Advertising)
```




```
attach(Advertising)
```

Question 2: Simple Regressions

Is there a relationship between sales and advertising media? To provide a preliminary analysis, run simple regressions of sales on each of the regressors (e.g. `lm(Sales~TV)`). Interpret the coefficients. Do all media contribute to sales?

```
fit1<-lm(Sales~TV)
summary(fit1)
```

```
##
## Call:
## lm(formula = Sales ~ TV)
##
## 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
```

A $1,000 increase in TV budget increases expected sales by 47 units

```
fit2<-lm(Sales~Radio)
summary(fit2)
```

```
##
## Call:
## lm(formula = Sales ~ Radio)
##
## 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
```

A $1,000 increase in Radio budget increases expected sales by 202 units

```
fit3<-lm(Sales~Newspaper)
summary(fit3)
```

```
##
## Call:
## lm(formula = Sales ~ Newspaper)
##
## 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
```

A $1,000 increase in Newspaper budget increases expected sales by 54 units All media seem to contribute to Sales, as their coefficient is statistically significant at 5%

Question 3: Multiple Regression

To provide a better answer run a multiple regression of sales on tv radio and newspaper. Interpret the slope estimates. Revisit the question of whether all media contribute to sales. How do you reconcile the results for the multiple and simple regressions for newspaper? How strong is the relationship between advertising and sales? Compute R-squared and discuss. Provide a 3D plot with model using just TV and Radio (use the library car and the command `scatter3d(Sales~TV+Radio)`)

```
fit4<-lm(Sales~TV+Radio+Newspaper)
summary(fit4)
```

```
##
## Call:
## lm(formula = Sales ~ TV + Radio + Newspaper)
##
## 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 ***
## Radio        0.188530   0.008611  21.893   <2e-16 ***
## Newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## 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
```

A $1,000 increase in TV budget increases expected sales by 45 units, holding Radio and Newspaper fixed. A $1,000 increase in Radio budget increases expected sales by 188 units, holding TV and Newspaper fixed. A $1,000 increase in Newspaper budget increases expected sales by 1 unit, holding TV and Radio fixed. TV and Radio contribute to Sales, but once we control for them, Newspaper does not contribute (its p-value is very large). Newspaper does not provide additional information to predict Sales, beyond what is already in Radio and TV. The relation is strong, in the sense that TV and Radio are able to explain a big proportion of the variability in Sales. The R2 is very high, 0.89.

```
library(car)
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
scatter3d(Sales~TV+Radio)
```

```
## Loading required namespace: rgl
```

Question 4: Model with Interactions

Is there a synergy among the advertising media? To see this run a multiple regression with an interaction between tv and radio (use `lm(Sales~TV*Radio)`. Does this model fit the data better?

```
fit5<-lm(Sales~TV*Radio)
summary(fit5)
```

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

The interaction term is quite significant, with a very small p-value, so there is a synergy effect. The fit is much better, with an R2 of 0.96.

Question 5: Optimize Sales

Your client has a budget of 300K. Based on previous results they should be divided between tv and radio (as newspaper was not significant). What is the optimal allocation based on the previous fitted model with interactions? Maximize the predicted sales by substituting radio=300-tv in the predicted sales and set to zero the first order condition. What are the optimal sales for the optimal combination? Compare with observed sales. Compute a confidence interval for the prediction. Hint: With the fitted model solve the optimization problem by hand to obtain the optimal values of TV and Radio. Then use the command `predict` in R.

This problem will be solved in class.

Part 2: Reading

Read the paper by Varian (2014), which is posted in Canvas. This is a nice paper that contains many ideas that we will be covered in class. Do not worry if you do not understand these ideas yet. The goal is that you will be able to understand them by the end of this course. After reading the paper answer the following general questions. What is the goal of Machine learning? What does Varian mean by "good out of sample predictions"? What is overfitting? What is model complexity? What is the training data?

This paper will be discussed in class, after Chapter 8.