Simple Regression Analysis

Xiaoqian Zhu

Abstract

In this report we extend the scope of the previous HW that apply computtaional tools to reproduce the main results displayed in section 3.2 (page 71 to 82) of the book *An Introduction to Statistical Learning* by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. Link to the book

Introduction

The overall goal is to provide advice on how to improve sales of the particular product. More specificcaly, the idea is to see whether there is an association between advertising on medium and sales, and if so, which medium would be the best to improve sales of the particular product. In this homework, we are specificaally looking at how the advertisement on TV, radio and newspaper affect sales of the particular product; in other words, we are seeing advertising on TV, newspaper and radio as explainary variables to explain the reponses to sales. I will first check the assiociation between each variable and sales; then, I am going to develop an multivariable regression model that can be used to estimate sales on the basis of advertising budget on TV, radio and newspaper using the advertising data set, which compiled by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani.

Data

In this homework, we are working with Advertising data set (Link for Advertising.csv), which consists of *Sales* (in thousands of units) of a particular 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*. In this homework, we are working on the three media and *sales* to find out weather they have an association.

Methodology

Looking at the data set, we are going study the relationship between *Sales* and a group of variables, including, *TV*, *Radio*, and *Newspaper*. We start our analysis by setting up null and alternative hypothesis. The null hypothesis is, H_0:$\beta_1 = \beta_2 = \beta 3 = 0$, that there is no relationship between TV, radio and newspaper advertising budget and sales; the alternative hypothesis is, H_1, that there exists one β_i that not equal to 0. To test the hypothesis, we apply a multiple linear regression model,$Sales = \beta_0 + \beta_1 TV + \beta_2 Radio + \beta_3 Newspaper + error$, on the data set, Advertising.csv, to get the estimates of β_0, β_1,β_2, and β_3, which are $\hat{\beta_0}$, $\hat{\beta_1}$, $\hat{\beta_2}$, and $\hat{\beta_3}$. In this homework, we will also use the least squares model to our data for the regression analysis like the previous one.

Results

We first apply a simple linear regression on each expaining variable tables with Sales. Table 1 shows results from the simple single regression of sales on TV.

```r
load('../data/regression.RData')
library("xtable")
table_TV <- xtable(regression_sum_TV,caption = 'Simple Regression of Sales on TV',digits = 4)
```

```r
row.names(table_TV) <- c('(Intercept)','TV')

print(table_TV, caption.placement = 'top',comment = getOption("xtable.comment", FALSE))
```

Table 2 shows the results from simple single regression of sales on radio.

```r
load('../data/regression.RData')
library("xtable")
table_Radio <- xtable(regression_sum_Radio,caption = 'Simple Regression of Sales on Radio',digits = 4)

row.names(table_Radio) <- c('(Intercept)','Radio')

print(table_Radio, caption.placement = 'top',comment = getOption("xtable.comment", FALSE))
```

Table 3 shows the results from simple single regression of sales on newspaper.

```r
load('../data/regression.RData')
library("xtable")
table_Newspaper <- xtable(regression_sum_Newspaper,caption = 'Simple Regression of Sales on Newspaper',d

row.names(table_Radio) <- c('(Intercept)','Newspaper')

print(table_Newspaper, caption.placement = 'top',comment = getOption("xtable.comment", FALSE))
```

Then, we apply the a multiple linear regression of sales on TV, radio and regression. Table 4 shows the results from this multiple linear regression.

```r
load('../data/regression.RData')
library("xtable")
table_multi <- xtable(regression_sum_multi,caption = 'Multiple Regression Table',digits = 4)

row.names(table_Radio) <- c('(Intercept)','TV', 'Radio', 'Newspaper')

print(table_multi, caption.placement = 'top',comment = getOption("xtable.comment", FALSE))
```

Moremover, in this homework, we also compute the correlation matrix among TV, radio, newspaper, and sales. Table 5 shows the correlation matrix for TV, radio, newspaper, and sales for the Advertising data.

```r
load('data/coorelation-matrix.RData')
library("xtable")
correlation_matrix = cor(advertising[sapply(advertising, is.numeric)])
cor_matrix <- xtable(correlation_matrix, caption = 'Correlation matrix for TV, radio, newspaper, and sal
print(table_multi, caption.placement = 'top',comment = getOption("xtable.comment", FALSE))
```

When we perform multiple linear regression, we usually are intrested in answering a few important questions.

1. Is at least one of the predictors useful in predicting the response?

2. Do all predictors help to explain the response, or is only a subset of the predictors useful?

3. How well does the model fit the data?

4. How accurate is the prediction?

we now address each of these questions in turn.

One: Is at least one of the predictors useful in predicting the response?

Recall that in the simple linear regression setting, in order to determine whether there is a relationship between the response and the predictor we can simply check whether $\beta 0 = 0$. Now, we are performing a multiple linear regression with 3 predictors, thus we need to ask whether all of the regression coefficients are 0. In this case, whether $\beta_1 = \beta_2 = \beta 3 = 0$. As in the simple linear regression setting, we use a hypothesis test to answer this question. We test the null hypothesis versus the alternative (which I specifically listed in introduction). Such hypothesis test is performed by computing the F-statistic. WHen there is no relationship between the response and predictors, one would expect the F-statistic to take a value close to 1. On the other hand, if alternative hypothesis is ture, then we expect F-statistic to be greater than 1.

```
library(xtable)
Quantity <- c('RSS', 'R2', 'F-stat')
Value <- c(regression_sum_multi$sigma,
           regression_sum_multi$adj.r.squared,
           regression_sum_multi$fstatistic[1])
results2 <- data.frame(Quantity, Value)

table2 <- xtable(results2, caption = 'More information about the least squares model for the multiple r

print(table2, caption.placement = 'top', comment = getOption("xtable.comment", FALSE))
```

The The F-statistic for the multiple linear regression model obtained by regressing *sales* onto *radio*, *TV*, and *newspaper* is shown in Table 6. In this example, the F-statistic = `load('../data/regression.RData'); r round(regression_sum_multi$fstatistic[[1]], digits = 0)`. Since this is far larger than 1, it provides compelling evidence against the null hypothesis. In other words, the large F-statistic suggests that at least one of the advertising media must be related to *sales*.

Two: Do all predictors help to explain the response, or is only a subset of the predictors useful?

It is possible that all of the predictors are associated with the response, but it is more often the case that the response is only realted to a subset of the predictors. Then, we would like to perform a variable selection that fit a single model involving each predictor to determine which predictors are associated with the response. Looking at each test's p-value, we are able to explain whether it is associated with *sales*. If the p-value is small, then we would be able to believe that there is a relationship between the single predictor with *sales*. On the other hand, if the p-value is large, then we have reason to believe that there is no relationship between two. Looking at Tables of single linear regressions, we can see that the estimated coefficient on TV is

Three: How well does the model fit the data?

Two of the most common numerical measures of model fit are RSE and R^2. An R^2 values close to 1 indicates that the model explains a large protion of the variance in the response variable. We saw in Table 6, for the Advertising data set, the model that uses all there advertising media to predict sales has an R^2 value of `load('../data/regression.RData'); round(regression_sum_multi$r.squared, digits = 4 )` This is very close to 1, which means that our explainatory varibales work well to explain changes in sales. Furthermore, RSE is the estimate of the standard deviation of error, and when it is small, it indicates a well-fitted model on data. In this multiple regression model, the RSE = `load('../data/regression.RData'); r round(regression_sum_multi$sigma, digits = 4)` . Thus, we can also tell from RSE that is a well-fitted model on advertising data set.

Four: How accurate is the prediction?

There are uncertainty associated with the estimated prediciton of sales, to answer how accurate is the prediction, we can simply look at the standard error of each estimated coefficients in Table 4. The standard of TV, Radio and newspaper are all very small, which means that we can be confidently sure that the estimation of explaintary variables are close to the true model. Furthermore, the confidence can also be shown by Rˆ2 like the previous question, that the Rˆ2 value is close to 1, which means that the prediction model is well-fitted.

Conclusions

To be honest, it is a very long homework, that we try to recreate the study that was done in chapter 3.2 of *An Introduction of Statistical Learning* by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. We first extend our analysis in previous homework in order to accomodate two more predictors, *Radio* and *Newspaper* by running two more simple linear regression. Later, we perform a multiple linear regression for better prediction on sales. From the results, we can conclude that the estimated coefficients of TV and *Radio* are statistically significant, while *Newspaper* seems to do not have a relationship towards sales, according to p-value. $1000 increase in TV advertising budget would lead to average increase in sales by 46 units, and $1000 increase in Radio advertisng budget would lead to average increase in sales by 189 units.Such multiple linear regression has a Rˆ2 close to 1 and a small RSE, thus, this model is well-fitted. The standard error of each variable shows that the estimation is close to the true model.