Multiple Regression Analysis

Jie Sun

October 14, 2016

Abstract

In this report we reproduce the analysis from Section 3.2 *Multiple Linear Regression* (chapter 3) of the book **An Introduction to Statistical Learning**.

Introduction

Given the `Advertising` data, the overall goal is to provide advice on how to improve sales of the particular product. More specifically, the idea is to determine whether there is an association between advertising and sales, and if so, develop an accurate model that can be used to predict sales on the basis of the predictor variables `TV`, `Radio` and `Newspaper` on which advertising budget has been spent.

In previous assignment `stat159-fall2016-hw02`, we came up with a simple linear regression model which predicts a response `Sales` on the basis of a single predictor variable `TV`. However, in practice we often have more than one predictor. A better approach is to extend the simple linear regression model so that it can directly accommodate multiple predictors. So this analysis focuses on fitting a multiple linear regression model to regress `Sales` onto `TV`, `Radio` and `Newspaper`.

Data

The Advertising data set 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`).

Methodology

To extend the simple regression model, we can give each predictor a separate slope coefficient in a single model. In other words, we're fitting the model:

$$Sales = \beta_0 + \beta_1 TV + \beta_2 radio + \beta_3 newspaper + \epsilon$$

To estimate the coefficients $\beta_0 ... \beta_3$, we fit the model via the least squares criterion.

Results

Running three separate simple linear regressions, we compute coefficient estimates of simple regression models: `Sales` on `TV`, `Sales` on `Radio`, and `Sales` on `Newspaper` in Table 1-3.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.0326	0.4578	15.36	0.0000
tv	0.0475	0.0027	17.67	0.0000

Table 1: Simple regression of sales on TV

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.3116	0.5629	16.54	0.0000
radio	0.2025	0.0204	9.92	0.0000

Table 2: Simple regression of sales on radio

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.3514	0.6214	19.88	0.0000
newspaper	0.0547	0.0166	3.30	0.0011

Table 3: Simple regression of sales on newspaper

Back to multiple regression, Table 4 displays the coefficient estimates of the least squares model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.9389	0.3119	9.42	0.0000
tv	0.0458	0.0014	32.81	0.0000
radio	0.1885	0.0086	21.89	0.0000
newspaper	-0.0010	0.0059	-0.18	0.8599

Table 4: Multiple regression of sales on TV, radio and newspaper

Consider the correlation matrix for the three predictor variables and response variable, displayed in Table 5. It shows that multiple regression suggests no relationship between `sales` and `newspaper` while the simple linear regression implies the opposite.

We also care about other information of the model like RSE, R^2 and F-statistic listed in Table 6.

Conclusions

Now we're done with the regression analysis, we can address some questions related to the fitness and accuracy of this model. The large F-statistic suggests that at least one of the advertising media must be related to sales. However, the p-values in Table 4 indicate that `TV` and `radio` are related to `sales`, but that there is no evidence that `newspaper` is associated with `sales`, in the presence of these two. Therefore, only a subset of the predictors is useful to explain the response.

In terms of the model itself, RSE and R^2 are both good measures. Given that R^2 is pretty close to 1, the model with three predictors fits most of the `Advertising` data. To examine the effect of each predictor, we need to take a closer look at R^2 of those pairwise simple regression models.

	TV	Radio	Newspaper	Sales
TV	1.0000	0.0548	0.0566	0.7822
Radio	0.0548	1.0000	0.3541	0.5762
Newspaper	0.0566	0.3541	1.0000	0.2283
Sales	0.7822	0.5762	0.2283	1.0000

Table 5: Correlation matrix for TV, radio, newspaper, and sales for the Advertising data

	Quantity	Value
1	RSE	1.69
2	R^2	0.897
3	F-statistic	570

Table 6: More information about the least squares model