Multiple Regression Analysis

Abby Vogel

October 14, 2016

Abstract

This paper uses both simple and multiple linear regression of Sales and advertising spending from a dataset of 200 markets. Advertising in this data set was through TV, Radio, and Newspaper and was measured in thousands of dollars spent. Utilizing the least squares method, the model was fit to the data and analysed. This model determined that both `TV` and `Radio` are important predictor variables in explaining differences in `Sales` between firms. `Newspaper` advertising did not show to be a predictor of `Sales` in the linear model. Further analysis of the model of just `Sales ~ TV + Radio` is needed to determine the most effective and efficient model of this data.

Introduction

The goal of this paper is to use linear regression to analyse how money spent on advertising and Sales are correlated for the purpose of improving marketing. By utilizing the tools of R, this project will create a linear model of the data and use it to better understand the relationship bettween advertisement spending and Sales. By utilizing the multiple linear regression, the effect of each form of adverstising can be assessed and firms can better utilize their advertising expediture.

With constant development of new forms of medium, companies have many more options in advertising. Current business theory provides the notion that spending on advertising over different mediums and with different subjects can drive sales. By isolating each form of adversiment in this period, we hope to determine if spending on TV, Radio, or Newspaper advertisements are correlated with higher overall sales.

Data

The data of this analysis comes from the text *An Introduction to Statistical Learning* authored by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. This data comes from 200 markets, with response variables of thousands of dollars spend on TV, Radio, and Newspaper advertisement, as well as the number of units sold (in thousands).

Methodology

Looking specifically at the TV, Radio, Newspaper and Sales data, we ran a linear model of the Sales (in thousands of units sold) onto each of the advertising mediums (in thousands of dollars spent).

We based our model on the simple linear model with an intercept and coefficient:

$$Sales = \beta_0 + \beta_1 X$$

To find the intercept and coefficient, the data was fit using `lm()` in R. This process utilized the least-squares regression method.

For the multiple regression, we fit the model using a linear model with and intercept and coefficient:

$$Sales = \beta_0 + \beta_1 TV + \beta_2 Radio + \beta_3 Newspaper$$

Results

Table 1: Simple Regression of Sales on TV

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.03	0.46	15.36	0.00
TV	0.05	0.00	17.67	0.00

Table 2: Simple Regression of Sales on Radio

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.31	0.56	16.54	0.00
Radio	0.20	0.02	9.92	0.00

Table 3: Simple Regression of Sales on Newspaper

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.35	0.62	19.88	0.00
Newspaper	0.05	0.02	3.30	0.00

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

A predictor is useful in predicting the response if $\beta_1 \neq 0$. For this model, if any of the coefficients is not equal to zero, then they are useful predictor variables. In the multiple regression (Table 4), both TV and Radio have $\beta_1 > 0$ so they are useful predictors in this model.

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

Not all predictors help explain the response in the multiple regression. Newspaper has a $\beta_1 = 0$ in the multiple regression, so it is not a useful predictor of sales in the multiple regression. This differs from the simple linear regression of Sales on Newspaper (Table 3), which indicates Newspaper as a useful predictor with $\beta_1 > 0$. This is only true of the simple model and when the other variables are added, Newspaper is no longer a predictor of sales. It is more appropriate to use a subset of the predictors to explain the response. A better model would be `lm(Sales ~ TV + Radio)`.

3. How well does the model fit the data?

The model fits the data well. The overall R^2 of this regression is 0.8972106 (Table 6), indicating that 89.7210638% of the variability of the data is accounted for in the model.

The RSE of this model is 1.6855104 (Table 6). This value shows how far any fitted value is from the true value in the model. Each prediction with be about 1.6855104 thousands of units away from the true value of sales. In comparison with the range of this data, this is an acceptable error but perhaps a different model would be a better fit on this data.

4. How accurate is the prediction?

This model has many sources of error, from reducible error, model bias, and irriducable error. A 95% confidence interval could be generated around sales to have an envelope that would cover the mean about 95% of the time it was generated.

Conclusions

The multiple regression model of `lm(Sales ~ TV + Radio + Newspaper)` is an improvement over the model `lm(Sales ~ TV)`. However, the variable `Newspaper` does not show to be a predictor in the level of Sales so it would be more appropriate to use the model `lm(Sales ~ TV + Radio)` in this multiple regression. Overall,

Table 4: Least Square Coefficient Estimates

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.94	0.31	9.42	0.00
TV	0.05	0.00	32.81	0.00
Radio	0.19	0.01	21.89	0.00
Newspaper	-0.00	0.01	-0.18	0.86

Table 5: Matrix of Correlation Coefficients

	TV	Radio	Newspaper	Sales
TV	1.00	0.05	0.06	0.78
Radio	0.05	1.00	0.35	0.58
Newspaper	0.06	0.35	1.00	0.23
Sales	0.78	0.58	0.23	1.00

the three variables account for about 90% of the variability in `Sales`, with `TV` as the strongest predictor of `Sales`. For a firm with a resonable budget, their sales figures would be most effected by increases in TV and Radio advertising expenditure. However, this is only true of firms like those in the sample, these results cannot be applied out of the scope of the data (in different cities, industries, etc.).

Table 6: More Least Squares Terms

Value	Quantity
RSE	1.69
R2	0.90
F-Stat	570.27