Multiple Linear Regression Analysis

Tina Huang

October 14, 2016

Abstract

This report will reproduce the findings concerning the relationship between TV, Radio, and Newspaper Budgets and Sales from the Advertising.csv dataset, from Chapter 3.2 of an *An Introduction to Statistical Learning* by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani.

Introduction

The purpose of this report is to improve overall sales by determining whether or not sales can be predicted by TV, Radio, and or Newspapers budgets. To do this, the relationship between these variables will be explored through simple linear regression, as well as multiple linear regression.

Data

The Advertising.csv dataset contains the TV, Radio, and Newspaper budgets in thousands of dollars, as well as sales in thousands of units, for 200 different markets. In this report, only the TV budgets and sales will be looked at from this dataset.

Methodology

We assume a roughly linear relationship between Sales and TV Budget, Sales and Radio Budget, and Sales and Newspaper Budget for the Advertising dataset, and so will use simple linear models to illustrate this relationships:

$$\text{Sales} = \beta_0 + \beta_1 \text{ TV}$$

$$\text{Sales} = \beta_0 + \beta_1 \text{ Radio}$$

$$\text{Sales} = \beta_0 + \beta_1 \text{ Newspaper}$$

β_0 and β_1 represent the intercept and slope of the equation, respectively. To estimate these unknown coefficients, we can minimize the least squares criterion.

For these simple linear regressions, first the summaries and histograms for the distributions of the variables are looked at individually. Then the coefficients are estimated, the standard error of these estimates is calculated, and then a t-test is performed to determine the significance of these values. In addition, the residual standard error and R-squared values can be calculated to determine how well the model fits the data.

We will also create multiple linear regression model of the form

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

to determine the effect of multiple predictors, TV, Radio, and Newspaper budget, on sales.

As with simple linear regression, the unknown coefficients are estimated using the least squares approach, in which we select values that minimize the residual sum of squares. A correlation matrix will also give us insight into the relationship between the variables.

To determine if at least one of TV, Radio, and Newspaper budget has an effect on sales, we will find the F-statistic of the model. Like before, to determine how well the model fits the data, the residual standard error and R-squared values will also be calculated.

Results

Table 1: Estimating the Regression Coefficients for Sales Onto TV

	Estimate	Std. Error	t value	Pr(>|t|)
TV	0.04754	0.002691	17.67	1.467e-42
(Intercept)	7.033	0.4578	15.36	1.406e-35

Table 2: Assessing the Fit of the Sales Onto TV Model

Observations	Residual Std. Error	R^2	Adjusted R^2
200	3.259	0.6119	0.6099

Table 3: Estimating the Regression Coefficients for Sales Onto Radio

	Estimate	Std. Error	t value	Pr(>|t|)
Radio	0.2025	0.02041	9.921	4.355e-19
(Intercept)	9.312	0.5629	16.54	3.561e-39

Table 4: Assessing the Fit of the Sales Onto Radio Model

Observations	Residual Std. Error	R^2	Adjusted R^2
200	4.275	0.332	0.3287

Table 5: Estimating the Regression Coefficients for Sales Onto Newspaper

	Estimate	Std. Error	t value	Pr(>|t|)
Newspaper	0.05469	0.01658	3.3	0.001148
(Intercept)	12.35	0.6214	19.88	4.714e-49

Table 6: Assessing the Fit of the Sales Onto Newspaper Model

Observations	Residual Std. Error	R^2	Adjusted R^2
200	5.092	0.05212	0.04733

The regression coefficients are estimated in Table 1 above, and we can see that β_0 and β_1 are estimated to be 7.0325935, 0.0475366 respectively. This means for an increase of $1000 in TV budget, we will see about an additional 47.5 units sold. We can also see that the p-value from performing the t-test is extremely low, indicating that this is a statistically significant value and that there is strong evidence that there is a relationship between TV Budget and Sales.

Looking at Table 2, we see that the model has an R-squared value of 0.6118751, which indicates a majority of the variability in Sales can be explained by TV Budget, and that the relationship between TV Budget and Sales is roughly linear. The residual standard error(RSE) value of 3.2586564 also supports these statements, as it is greater than 1.

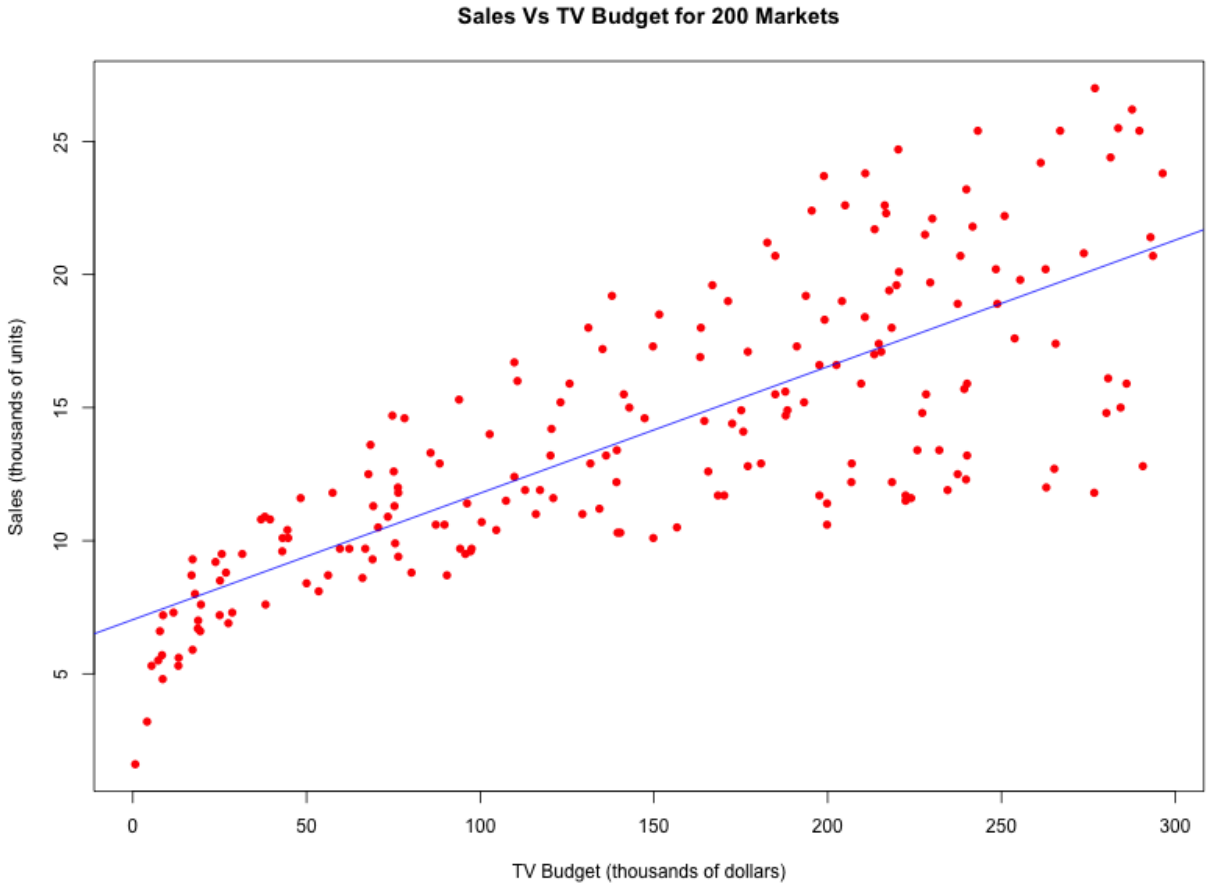

Figure 1: Scatterplot of Sales Vs TV Budget with Fitted Regression Line

Looking at Figure 1, we can see that the fitted regression line roughly encompasses most of the data points, however for low TV budgets, the line does not fit the data as well. The values also tend to fan out as TV budget increases, so while the relationship between the two variables seems to be approximately linear, there may be other non-linear models that could provide a better fit.

Conclusions

Based on our results, we see a very low p-value after performing a t-test to determine whether there's a relationship between Sales and TV Budget, which indicates that there is strong evidence that there is a relationship. The R-squared and RSE values indicate that there is a roughly linear relationship between these two variables. The scatterplot supports this claim, also giving us additional information that the fitted

model is not as accurate for low TV budgets. Therefore we can see that we can use the model to roughly predict Sales based on TV Budget, as long as TV Budget is above a certain value (approximately at least 10 thousand dollars).