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 for the Sales onto TV model 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 47.5366404 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 Sales onto TV 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 not too large.

Similarly for the Sales onto Radio model, in Table 3, we can see that β_0 and β_1 are estimated to be 9.3116381, 0.2024958 respectively. This means for an increase of $1000 in Radio budget, we will see about an additional 202.4957834 units sold. We can also see that the p-value from performing the t-test is also extremely low, indicating that this is a statistically significant value and that there is strong evidence that there is a relationship between Radio Budget and Sales.

Looking at Table 4, we see that the Sales onto Radio model has an R-squared value of 0.3320325, which indicates some of the variability in Sales can be explained by Radio Budget. The residual standard error (RSE) value of 4.2749444 also supports this statement, as it is not too large but also not too small either.

Finally for the Sales onto Newspaper model, in Table 5, we can see that β_0 and β_1 are estimated to be 12.3514071, 0.0546931 respectively. This means for an increase of $1000 in Newspaper budget, we will see about an additional 54.6930985 units sold. We can also see that the p-value from performing the t-test is also low, indicating that this is a statistically significant value and that there is evidence that there is a relationship between Newspaper Budget and Sales.

Looking at Table 6, we see that the Sales onto Newspaper model has an R-squared value of 0.0521204, which indicates almost none of the variability in Sales can be explained by Newspaper Budget. The residual standard error (RSE) value of 5.0924804 also supports this statement, as it is not very small.

Table 7: Least Squares Regression Coefficient Estimates for Sales Onto TV, Radio, and Newspaper

	Estimate	Std. Error	t value	Pr(>|t|)
TV	0.04576	0.001395	32.81	1.51e-81
Radio	0.1885	0.008611	21.89	1.505e-54
Newspaper	-0.001037	0.005871	-0.1767	0.8599
(Intercept)	2.939	0.3119	9.422	1.267e-17

Table 8: Assessing the Fit of the Multiple Regression Model

Observations	Residual Std. Error	R^2	Adjusted R^2
200	1.686	0.8972	0.8956

In Table 8, we can see that the multiple linear regression model seems to fit the data well as the R-squared value is 0.8972106, which indicates most of the variability in Sales can be explained by these three budgets. In addition, the residual standard error is quite low, with a value of 1.6855104, which also supports this statement.

We also calculate the F-statistic using equation 3.23 in the text, and obtain the value 570.2707037. Since this is a very high value, this supports the above statement as we can conclude that there is strong evidence that there is a relationship between Sales and at least one of these three budgets.

Looking at Table 7, we are better able to see the individual effects of each of the budgets on sales. As with the simple linear regression, both the coefficients for TV and Radio budgets have extremely low p-values, so there is strong evidence that there is a relationship between these budgets and sales. However with the multiple linear regression, we now see that the coefficient for Newspaper has a high p-value of 0.8599151, indicating that even though the simple linear regression found Newspaper budget to be statistically significant, the multiple linear regression model finds strong evidence for us to accept the null hypothesis that there is no relationship between Newspaper budget and Sales.

To explain why the simple linear regression found that there was strong evidence for a relationship between Newspaper budget and Sales but the multiple linear regression is finding the opposite, we can take a look at the correlation matrix for these four variables.

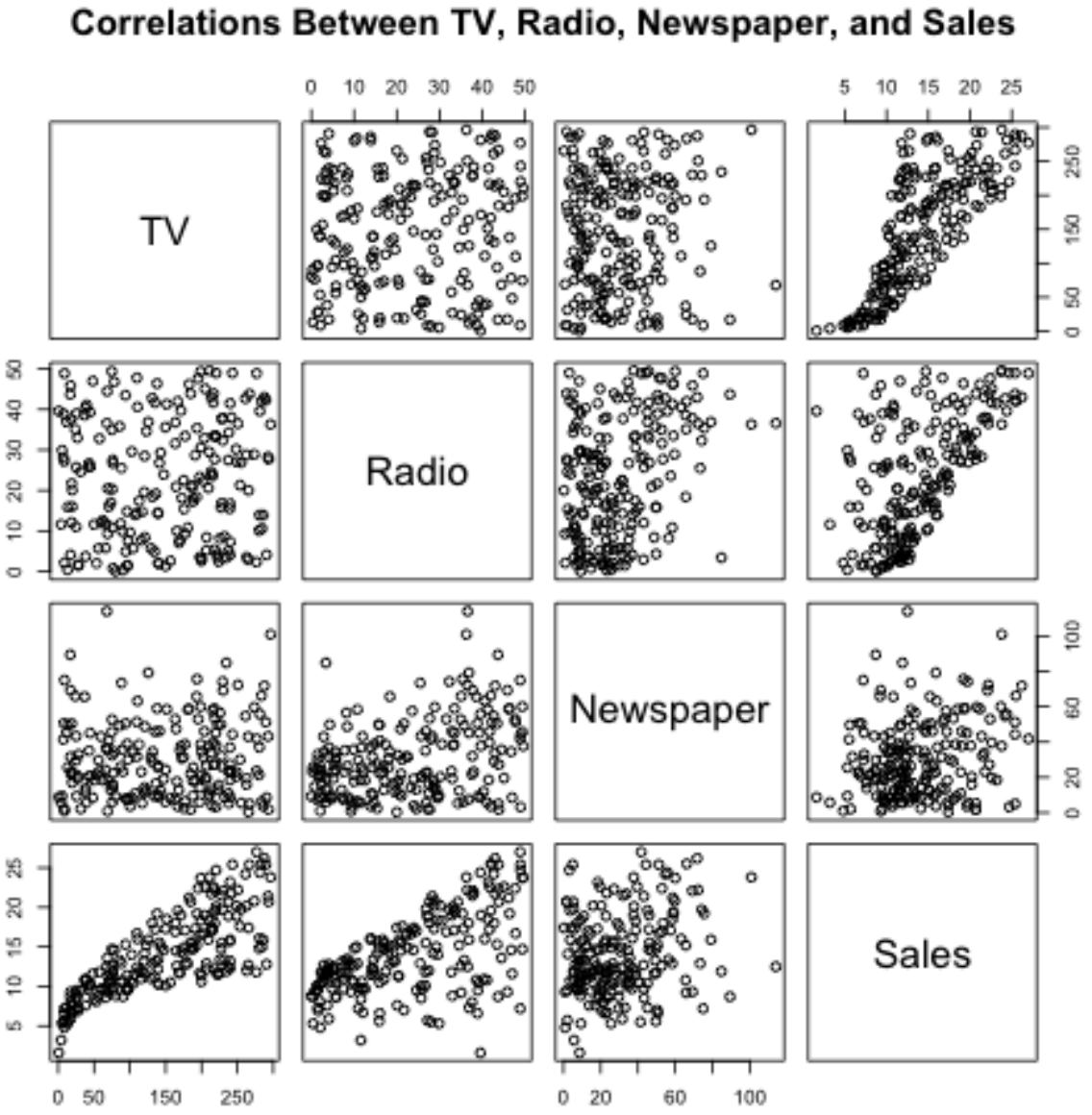

Figure 1: Correlation Matrix of Sales, TV, Radio, and Newspaper

Looking at Figure 1, we can easily see that TV and Sales are correlated, and that Radio and Sales are correlated as well. However, Newspaper and Sales do not seem to have a clear strong relationship. We

can also see though that Newspaper and Radio seems to have slight relationship, albeit a weak one. Since they are somewhat correlated, that means that an increase in Radio budget also tends to have an increase in Newspaper budget. We already know that Radio budget does have an effect on sales, so that's why it appeared that Newspaper budget had an effect on sales, while in reality, individually Newspaper budget and Sales have almost no correlation.

Conclusions

Based on our results, using the multiple linear regression F-statistic, we could see that at least one of TV, Radio, and Newspaper budgets are useful in predicting the response, Sales. Looking at the p-values for the coefficient estimates for these predictors, we see that two of our predictors, TV and Radio budget, are useful in predicting Sales. Newspaper budget, on the other hand, has a very low correlation with Sales. The low residual standard error and high R-squared value tell us that the multiple linear regression model is a good fit for our data. Since it is a good fit, our prediction for Sales will be relatively accurate given TV, Radio and Newspaper budgets.