Abstract

The purpose of the report is to extend the scope of the previous one. In this report, we want to carry out a multiple linear regression with predictor variables TV, Radio, Newspaper, and the response variable Sales. We want to reproduce the analysis in section 3.2: Multiple Linear Regression of the book “An Introduction to Statistical Learning” by Gareth James, Deniela Witten, Trevor Hastie and Robert Tibshirani.

Introduction

In this report, I will try to find out the relationship between sales and the other three factors, including the advertisement budget on TV, Radio and Newspaper. By conducting the multiple linear regression, we are able to find out the association between the predictor variables and the response variables so that we can provide advice on how to improve the sales.

Data

We use the advertising data set to conduct the multiple linear regression. The advertising data set can be downloaded from http://www-bcf.usc.edu/~gareth/ISL/Advertising.csv. The data set includes the advertising budget (in thousands of dollars) for TV, Radio, Newspaper and the Sales (in thousands of units) of a certain product in 200 different markets. In this report, we want to use all of the four variables in the data set to find if there’s a multiple linear relationship between them.

Methodology

To analyze the relationship between the four variables, I used the multiple linear regression model: $Sales = \beta_0 + \beta_1 * TV + \beta_2 * Radio + \beta_3 * Newspaper + e$. In this model, β_0 represents the intercept and β_1, β_2, and β_3 represent the slopes of the three predictor variables. I used the ordinary least squares estimation to estimate the coefficients and to fit a plane for the data set.

Results

	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 Linear Regression on TV and Sales

From table 1, we can see that the estimated slope is 0.0475. We can interpret the statistics as for every addtional one thousand dollars spent on TV budget, the sales will increase for 47.5 units. From the table, we can also see that the t-values for the estimated slope and estimated intercept are large while the p-values for them are small.

	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 Linear Regression on Radio and Sales

From table 2, we can see that the estimated slope is 0.2025. We can interpret the statistics as for every addtional one thousand dollars spent on Radio budget, the sales will increase for 202.5 units. From the table, we can also see that the p-values are small.

	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 Linear Regression on Newspaper and Sales

From table 3, we can see that the estimated slope is 0.0547. We can interpret the statistics as for every addtional one thousand dollars spent on Newspaper budget, the sales will increase for 54.7 units. From the table, we can also see that the p-values are small.

	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 Linear Regression

From table 4, we can interpret the slopes of TV and Radio as fixing all other factors constant, sales is predicted to increase by 45.8 units and 188.5 units respectively for every thousand dollar increase in the budget of TV and Radio. We can see that the t value for intercept, TV and Radio are large and p values for them are small, which indicates that the advertisement budget on TV and Radio has a linear regression relationship with Sales. However, from the table we can see that the p value for Newspaper is quite large and the t value is very small, which implies that Newspaper may not have a linear relationship with Sales. We want to look into more statistics to understand the results better.

	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

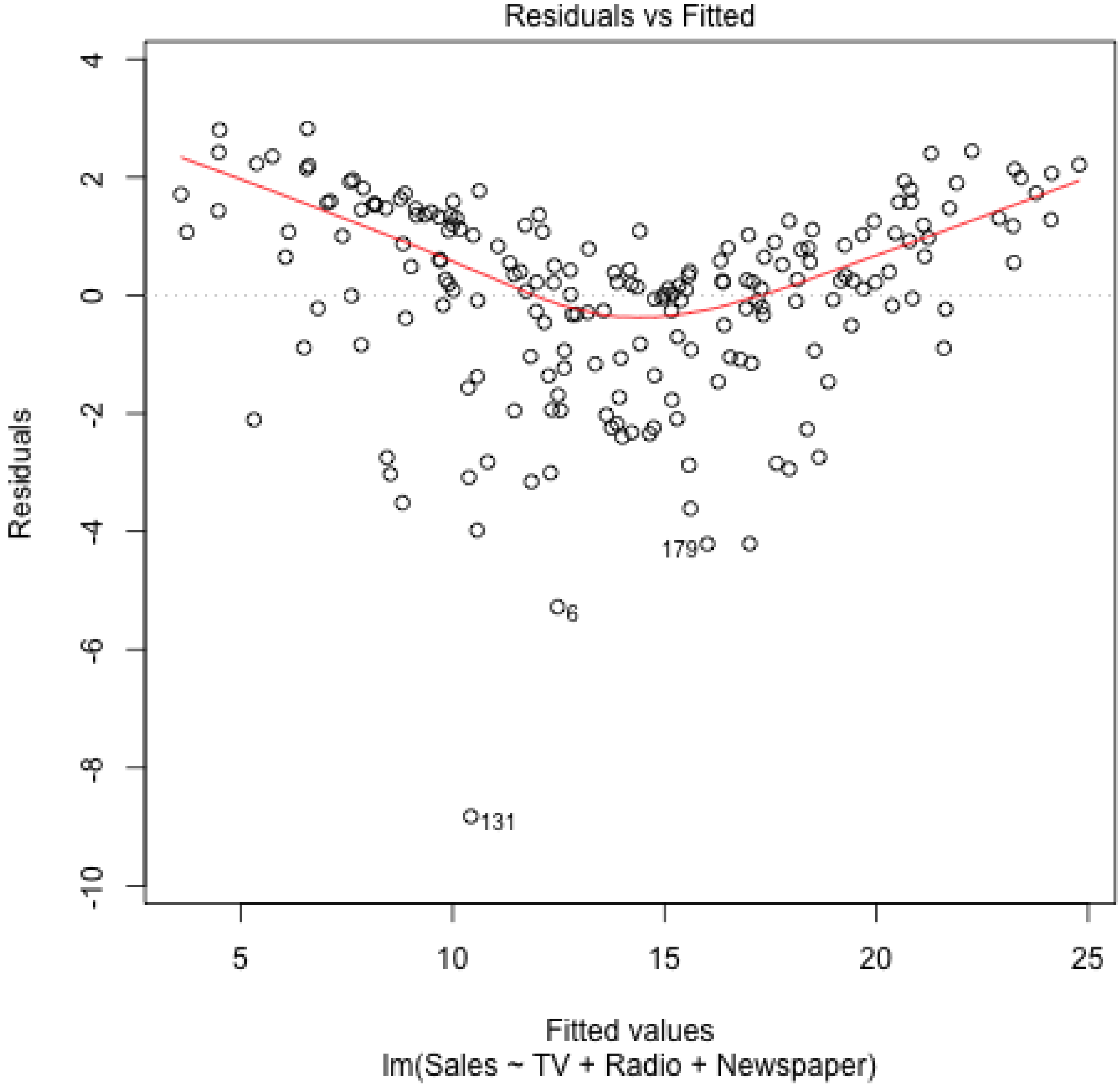
Table 5: Correlations Matrix

From table 5, we can see that there is weak correlation between sales and newspaper but strong correlation between sales and tv and sales and radio.

From the table 6, we can see that the r squared is large, which indicates a strong multiple linear relationship between the four factors. F statistic is very large, which furhter justifies that there's a strong multiple linear relationship between the four factors.

	Quantity	Value
1	RSS	13.62
2	R-square	0.90
3	F-Statistics	570.27

Table 6: Regression Quality Statistics



From the graph, we can see that the residuals are not quite random. The residuals are not randomly distributed around 0 and they seems to have a trend. We can interpret it as multiple linear regression model may not be the best model to fit the data set.

Conclusion

We can concludes that there's a linear relationship between tv and sales and radio and sales, but the relationship between newspaper and sales is weak. Overall, when fitting a multiple linear regression model to the data set, the model may not be the best model. We can try to remove the newspaper factor or fit some other models to find the best fit for the data set.