Regression Analysis

Elly Wang

10/13/2016

Abstract

In sections 3.1 and 3.2 of the book **An Introduction to Statistical Learning**, we were introduced to the concepts of simple and multiple linear regression. Specifically, we used the data set for advertising and performed simple and multiple linear regressions of Sales on TV, Radio, and Newspaper budgets, individually and collectively. In this report we will reproduce the graphs, regressions, and analysis presented in the book in a reproducible manner.

Introduction

The purpose of advertising for a certain good is to hopefully to increase its sales. With the technologies available today, there are many possible channels for advertisements to reach their audience. In this report, we will examine the relationships between the budget alloted for TV, Radio, and Newspaper advertisements and Sales of a particular good, and see if we can determine whether there is any relationships bewtween these budgets and the sales of the product.

Data

The advertising data used in the book consists of **Sales** (in thousands of units) of a particular prouduct in 200 different markets and advertising budgets (in thousands of dollars) for the product in each of those markets. In particular, the advertising budgets were for **TV**, **Radio**, and **Newspaper**.

Methodology

In this paper, we will examine relationships between:

- Sales (in thousands of units) and TV budget (in thousands of dollars),
- Sales (in thousands of units) and Radio budget (in thousands of dollars),
- Sales (in thousands of units) and Newspaper budget (in thousands of dollars),
- Sales (in thousands of units) and TV, Radio, and Newspaper budgets (in thousands of dollars)

In particular, our models for the simple linear regression will be:

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

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

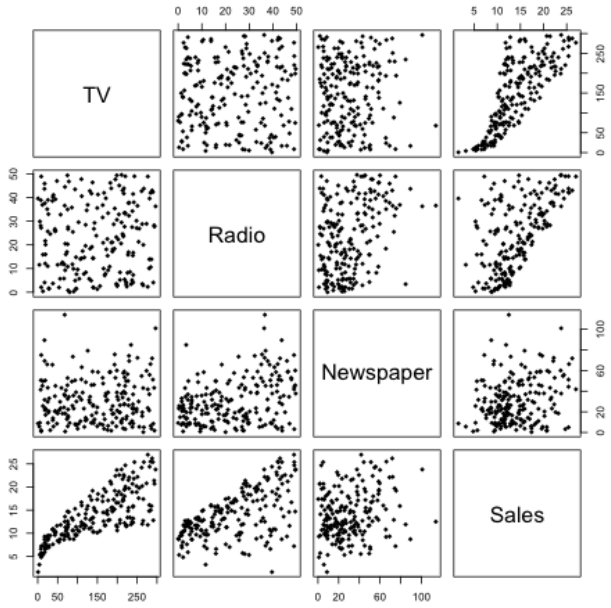
$$\text{Sales} = \beta_0 + \beta_1 * \text{Newspaper}$$

and our multiple linear regression model will be:

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

To esimate the coefficients for β_0, β_1, β_2, β_3, we will perform the ordinary least squares regression in R.

Results



Looking at the scatterplot matrix, we roughly get a sense of how each factor is related to sales. Focusing on the last row, we see that some general upward sloping trends for Sales on TV and Sales on Radio. In contrast, the scatterplot for Sales on Newspaper do not have a definte positive trend like the previous two plots do.

To get a better sense of how each factor relate to sales, we ran OLS regression on each pair of variables, and the estimates of the coefficients are shown below in Tables 1, 2, 3, and 4.

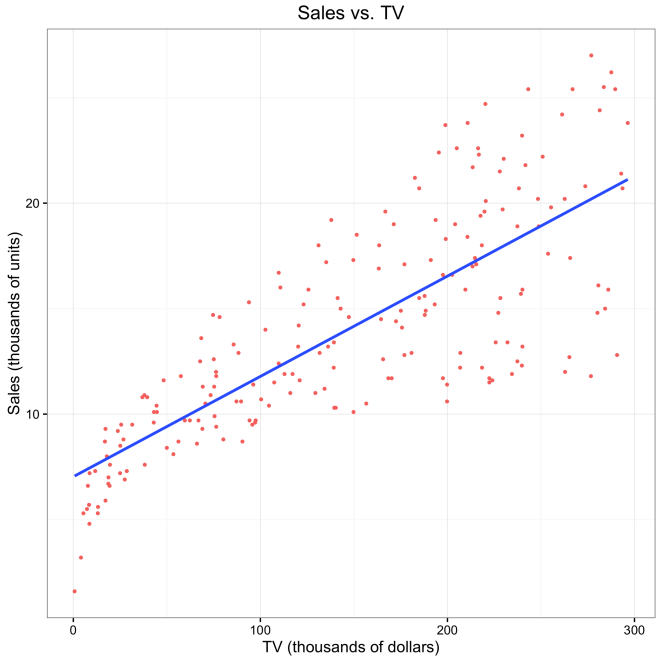
Sales and TV

	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

From table 1, we can see that for every thousand dollar increase in the budget for TV, sales is predicted to increase by 47.5366404 units. Since the standard error for the estimate of β_1 is quite low and the t-statistic is quite high, we can say that our estimate is signficiantly different from zero. The estimate of β_0 tells that that even when there are no budget alloted to TV, we'd expect the sales to be at around 7.0325935 thousands of units.

The scatterplot of Sales on TV with the regression lines fitted is shown below.



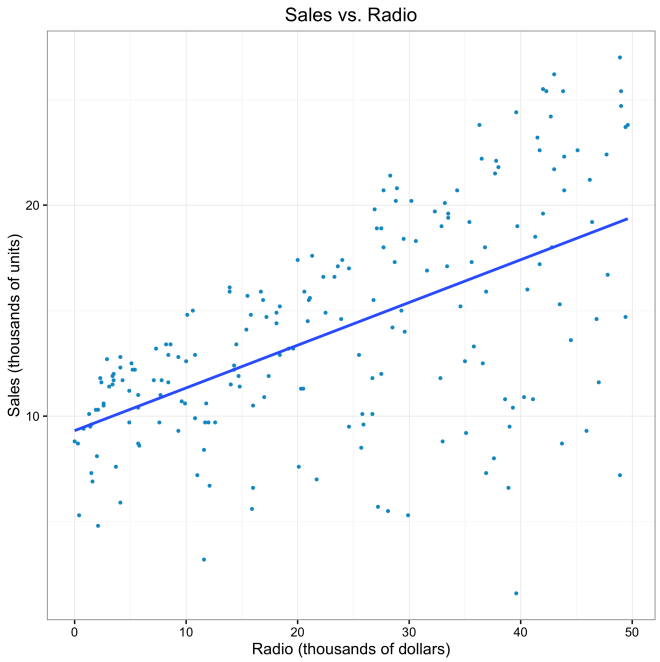
Sales and Radio

	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

From table 2, we can see that for every thousand dollar increase in the budget for Radio, sales is predicted to increase by 202.4957834 units. Since the standard error for the estimate of β_1 is quite low and the t-statistic is quite high, we can say that our estimate is signficiantly different from zero. The estimate of β_0 tells that that even when there are no budget alloted to Radio, we'd expect the sales to be at around 9.3116381 thousands of units.

The scatterplot of Sales on Radio with the regression lines fitted is shown below.



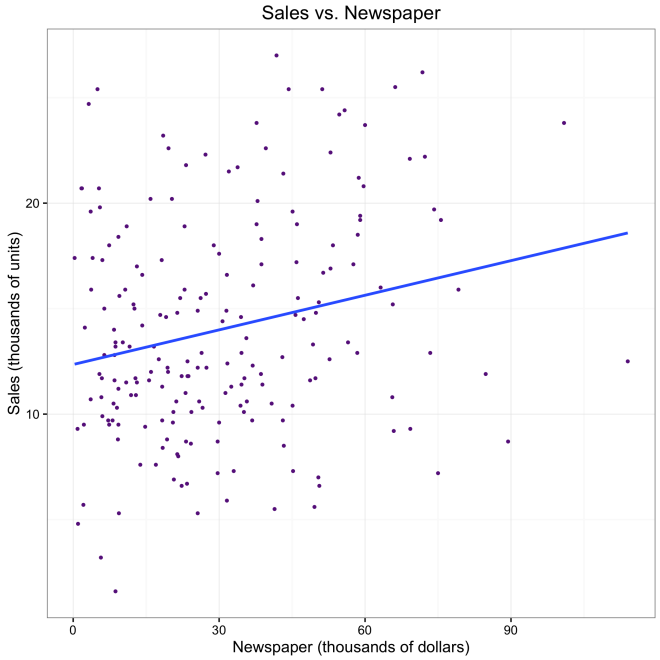
Sales and Newspaper

	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

From table 3, we can see that for every thousand dollar increase in the budget for Newsapper, sales is predicted to increase by 54.6930985 units. Since the standard error for the estimate of β_1 fairly quite low and the t-statistic is greater than 2, we can say that our estimate is signficiantly different from zero. The estimate of β_0 tells that that even when there are no budget alloted to Newspaper, we'd expect the sales to be at around 12.3514071 thousands of units.

The scatterplot of Sales on Newspaper with the regression lines fitted is shown below. Although initially we saw no obvious linear trend in the scatter, from simple linear regression, we see that is a weak linear relationship between newsapaper budget and sales.



Sales and all three factors

	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 see that for every thousand dollar increase in the budget for TV, while holding all other factors constant, sales is predicted to increase by 45.7646455 units. For every thousand dollar increase in the budget for Radio, while holding all other factors constant, sales is predicted to increase by 188.5300169 units. Lastly, for every thousand dollar increase in the budget for Newspaper, while holding all other factors constant, sales is predicted to increase by -1.037493 units.

Since the standard error for the estimate of β_1 is quite low and the t-statistic is quite high, we can say that our estimate is signficiantly different from zero. The estimate of β_0 tells that that even when there are no budget alloted to TV, we'd expect the sales to be at around 12.3514071 thousands of units.

	TV	Radio	Newspaper	Sales
TV	1	0.0548	0.0566	0.7822
Radio		1	0.3541	0.5762
Newspaper			1	0.2283
Sales				1

Table 5: Correlation Matrix of Advertising Data

Regarding the regression quality and the fitness of our multiple linear regression model, we reference the following table (Table 6).

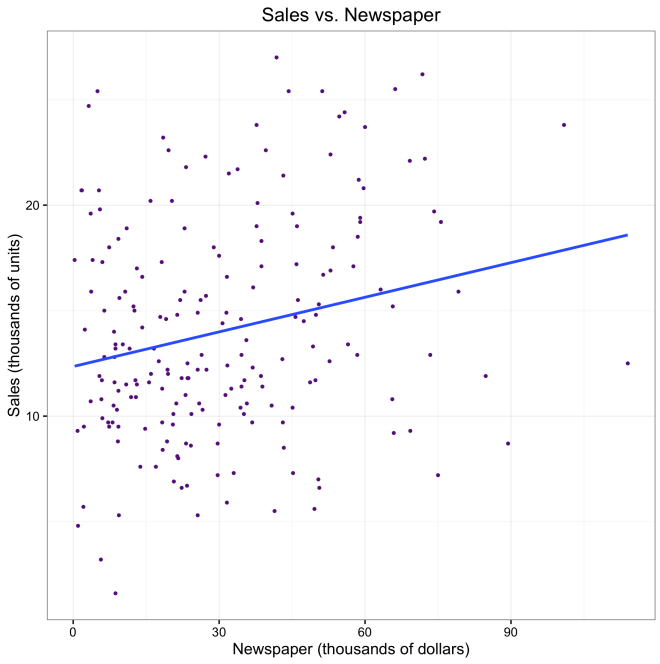
In this table, we see that the sum of the residuals squared is 1.69, which is a fairly low number for RSS. In doing OLS, we tried to minimize the sum of residuals squared because RSS is directly related to the RSE,

Quantity	Value
Residual standard error	1.69
R Squared	0.90
F-Statistics	570.27

Table 6: Regression Quality Indicies

which is a measure of lack of fit.
Another way to measure the fitness of a regression is through R^2 From the table, R^2 is 0.61, which means that 61% of the changes in sales is predicted by the changes in TV budget. This number is not considered too high, as 0.99 means close to perfect fit, but it's also not too bad for a fitness test. The high F-statistic in the table also tells us that the estimated coefficents in this regression is significantly different from zero.

The scatterplot of Sales on TV with the regression lines fitted is shown below in Figure 1.



Conclusions

Following the simple linear regression presented in section 3.1 of **An Introduction to Statistical Learning**, we were able to reproduce the graphs, regression model, and arrive at the same results and conclusions.

From the regression, we can see that the linear model we produced using OLS had a fairly good fit. From the siginficance of the regression coefficients and the R^2 value, we can conclude that Sales of the particular good is positively related to the budget alloted for TV advertisment. Although we cannot conclude that its a causal relationship, we can say that these two factors are positively correlated.

Lastly, specifically in the production of this report, we utilized git, github, R, and Makefile to create a streamline workflow that is easily reproducible.