Simple Regression Analysis

Austin Carango

October 31, 2016

Abstract

This report reproduces a simple linear regression model from chapter 3.1 of An Introduction to Statistical Learning with Applications in R by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. This model attempts to predict sales from TV advertising budget. Both variables are in the thousands.

1 Introduction

We would like to know how spending on TV advertising affects the sales of a particular product. Answering this question can help a company develop a marketing strategy to increase sales.

We will accomplish this by assuming there is a linear relationship between sales and money spent on advertising. To be more specific, we will asume that sales are a linear function of TV advertising and we will fit a linear model so as to attempt to predict how many sales a particular ad campaign will generate.

2 Data

The raw data for this report is contained in 'Advertising.csv', which has 200 observations of 5 variables. Look at the readme if you would like a URL to the data. The two variables we are interested in are 'TV' and 'Sales'. 'TV' is the advertising budget for television ads in thousands, and 'Sales' is a particular product's sales in thousands in 200 markets.

3 Methodology

For this linear regression, we treat 'Sales' as the response variable. This means 'Sales' will depend on a function of some other variable/s. 'TV' is the explanatory variable, meaning its value determines the value of 'Sales' via a function. We use the simple linear model below to estimate the relationship between 'Sales' and 'TV'.

$$Sales = \beta_0 + \beta_1 * TV \tag{1}$$

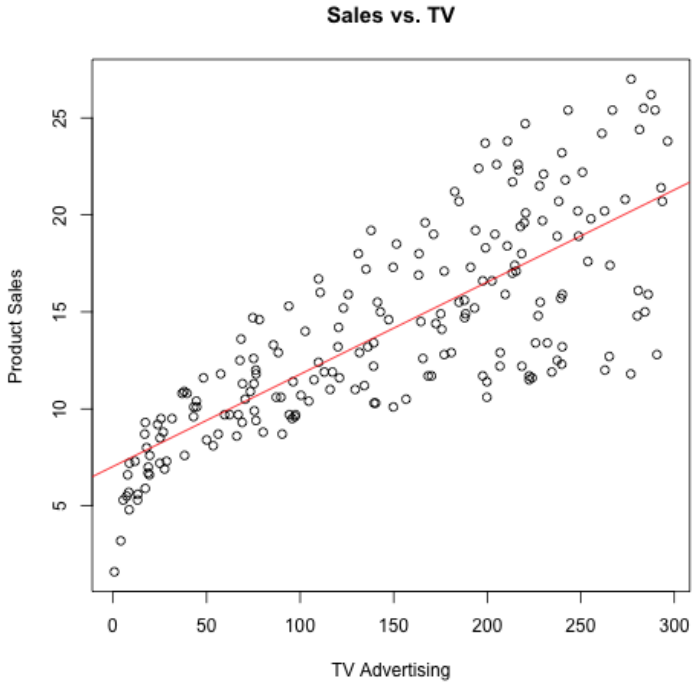
We have 2 beta coefficients which are the slope and intercept of the linear model. It is also evident that 'Sales' is the response variable and 'TV' is the explanatory variable.

In order to fit a line to this data we must minimize the least squared error. In other words we must find a line that has the least verticel distance between its points and the actual data points. We accomplish this by minimizing the expression below.

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \tag{2}$$

4 Results

The result of minimizing the least squared error gives us the line in figure 1. It is evident that the line is positively sloped, suggesting there is a positive correlation between 'Sales' and 'TV'. However, this scatterplot also appears to be heteroskedastic, with variance of 'Sales' increasing as 'TV' increases. Without homoskedasticity, the Gauss-Markov theorem does not apply, and thus the OLS coefficient estimators we obtain are not the best linear unbiased estimators (although they are still unbiased). Furthermore, this heteroskedasticity causes the standard errors of the coefficients to be biased, meaning the results of hypothesis tests have a higher risk of being incorrect. Let us further consider the model.



In the table below we have the estimates for the coefficients in the model, their standard errors, and numbers pertaining to hypothesis testing.

We see that the coefficient β_1 is estimated to be 0.05, meaning the model has determined there to be a mildly positive relationship between 'TV' and 'Sales'. We already noticed this in the graph.

The t test performed on each coefficient tests the null hypothesis that the coefficient equals 0 versus the alternate hypothesis that the coefficient is nonzero. We have very small p values in this model (both 0.00), suggesting that there is indeed some sort of relationship at play, whether or not the simple model is an extremely accurate measure of that relationship. Remember that we have heteroskedastic data, however, meaning we should not be fully confident in this test.

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

This second table contains some more useful numbers. R2 measures the proportion of total variability accounted for by the model. It takes on a value between 0 and 1, so in this case we see that the model accounted for a moderate amount of variability. We also have the residual standard error, which estimates the amount that our model is off from the actual data. The value we have is 3.26 which is also somewhat moderate for the data in question.

Quantity	Value
Residual SE	3.26
R2	0.61
F-stat	312.14

5 Conclusion

In conclusion, the simple model used in this report to describe the relationship between 'Sales' and 'TV' does a good job of showing us the intuitive conclusion we get from looking at the scatterplot of the data; there is a mild positive relationship between the two. What this model lacks is a way to account for the fact that the data is heteroskedastic. In order to make reliable inferences and have a more precise understanding of the relationship at play, a more complicated linear model is necessary.