Simple Regression Analysis

Thomas Sun

October 7, 2016

Abstract

This report attempts to reproduce the results found in Chapter 3 of the book **An Introduction to Statistical Learning**. In this chapter, a regression analysis is run on the *Advertising* dataset, containing data on sales and advertising budget for a particular product. Using a simple linear regression model, I find the same estimates of the coefficients, obtain the same quality index results, and similar looking plots as the ones contained in the book.

Introduction

The goal is to determine whether there is an effect of advertising on sales, ideally to increase product sales. Specifically, if increasing spending on certain mediums of advertising has a relationship with the amount of sales on a product. The chapter mainly considers one medium, `TV`, and fits a regression model to it with `Sales`. It finds that there is a strong positive relationship between `TV` and `Sales` and the data points fit the regression line closely.

Data

Data was obtained by downloading the *Advertising* dataset available on the textbook's website. It contains data of the size of advertising budget for `TV`, `Radio`, and `Newspaper` (in thousands of dollars) for a product in 200 different markets, in addition to the number of sales (in thousands of units) for the product in each market. We specifically are interested in the data for `TV` and `Sales`.

Methodology

In order to estimate the relationship between `TV` and `Sales`, a simple linear model is used.

$Sales = \beta 0 + \beta 1 * TV$

Where $\beta 0$ and $\beta 1$ are the regression coefficients, intercept and slope, respectively. To find an estimate for these coefficients, we use the ordinary least squares method to fit the model. OLS regression was run through RStudio, where the regression coefficients and quality indices (mean squared error, R-squared, F-statistic) are calculated. We also plot `TV` against `Sales` to replicate the scatterplot in the chapter.

Results

The estimated coefficients were calculated and found on the following table.

Table 1: Regression Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.033	0.458	15.360	0
TV	0.048	0.003	17.668	0

Both estimators have low standard deviation and p-value of practially zero, suggesting these calculated values are very statistically significant. In order to assess the fit of the data with the regression line, the following measurements were calculated in the table below.

Table 2: Regression Quality Indices

Quantity	Value
MSE	3.25865636865046
R-squared	0.611875050850071
F-Stat	312.144994372713

The mean squared error is relatively low while the F-Statistic is relatively large, suggesting the regression estimations are high quality. The scatterplot below is generated to visually interepret the data. A fitted regression line is added to compare the data points to the sample regression line. The data points seem to follow along the line quite closely with few outliers.

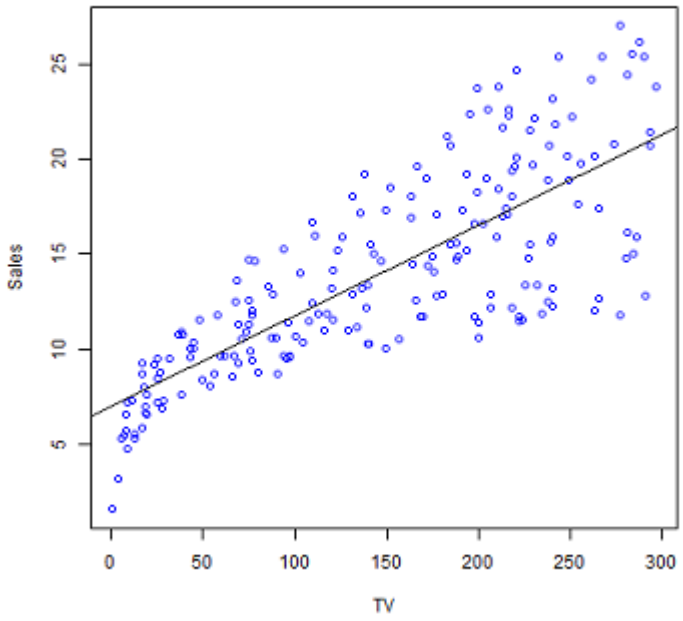

Figure 1: Scatterplot of `TV` and `Sales` with fitted regression line

Conclusions

Using a simple linear regression, I successfully managed to reproduce the regression analysis and results found in the textbook. The values for the estimated coefficients and quality indices are exactly the same. These results suggest that increasing TV advertising budget means the number of sales for the product will also increase. The regression coefficients appear they can be useful in predicting sales for a given amount of TV advertising. According to the regression line, for every additional unit of TV advertising, sales of the product will increase by .05 units.

However, based on the scatterplot, points seem to increase in variance as sales increases. This violates the simple linear regression assumption that the variables are homoskedastic, which weakens the efficiency of the

estimators. This means that variance in the estimators are likely to be larger and the calculated standard errors are likely biased. A more robust analysis can be performed to account for heteroskedasticity.

Additionally, when TV advertising is near zero, the data points seems to fall below the regression line. A non-linear model, such as a square root curve, may fit better.