Abstract

The following paper reproduces a simple linear regression model used in *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. The regression model pertains to predicting sales (in the thousands) from TV advertising budget (also in the thousands). It is features in chapter 3.1 of the book.

Introduction

Given a data set containing information pertaining to both sales and advertising budget, how could one develop a marketing plan that would result in higher product sales?

To achieve this goal, the following paper will display how improving sales via linear regression can be implemented. We will attempt to model the sales of a particular product as a function of advertising budget for three types of media: TV, radio, and newspaper. Because we are modelling this relationship via linear regression, thus inherently assuming that a linear relationship exists between our dependent variables and independent variable.

Data

The data set utilized in this paper, **Advertising.csv**, contains 200 observations of 5 variables. Four of the variables are of general interest: **Sales**, **TV**, **Radio**, and **Newspaper**. The fifth variable, **X**, is an index.

This dataset is free to access online at the following url [URL]. Alternatively, you can visit the author's GitHub repository to access it as well.

For this paper, we will explicitly use two variables: **Sales** and **TV**. **Sales** refers to the number (in thousands) of a particular product's sales in 200 different markets. **TV** refers to the advertising budget (in thousands) allocated to TV.

In our analysis, we will treat **Sales** as the dependent variable and **TV** as the independent variable.

Methodology

As stated, our approach for determining the association between **Sales** and **TV** will be linear regression. In particular, we will use simple linear regression: predicting **Y** (a continuous, quantitative response) on the basis of single predictor variable, **X**. We'll assume the following linear relationship, regressing Y onto X:

$$Y = \beta_0 + \beta_1 X$$

In our case, the model appears as follows:

$$sales = \beta_0 + \beta_1 * TV$$

Our *beta* parameters are unkown constants that represent the *slope* and *intercept* of our linear model. These parameters are known as the model *coefficients*. By fitting a linear model onto our training data, we obtain coefficient estimates that we use in the aforementioned linear model. Our new linear model, then, looks as follows.

We'll use this linear model to predict future values of our response (**Sales**).

Estimating the Coefficients

In order to obtain coefficient estimates, we need trainin data. Assuming we have a total of n observations, this data comes in the following form:

$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$$

Or, in our case with the *Advertising* data set:

$$(TV_1, Sales_1), (TV_2, Sales_2), ..., (TV_n, Sales_n)$$

Given a relevant training set, our goal is to obtain coefficient estimates for our *Beta* values such that our linear model fits the data as closely as possible. Put more succintly, we want to find the intercept and slope of our model such that the line resulting from our linear equation is as close as possible to our $n = 200$ points of data. To do this, we'll use the approach of minimizing the least squared error.

How do we do this? It's actually pretty simple: to get our line as close as possible to each dataset, we simply minimize the sum of the distances between our predicted values and the actual values. This process can be achieved by finding the values for our *beta* terms that minimize the following quantity:

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

Visually, this process looks as follows:

Determining a relationship between *X* and *Y*

Our obtained standard errors can be used to perform a hypothesis test, investigating as to whether or not a relationship exists between our dependent and independent variables. The hypotheses of this test generally assume the following structure:

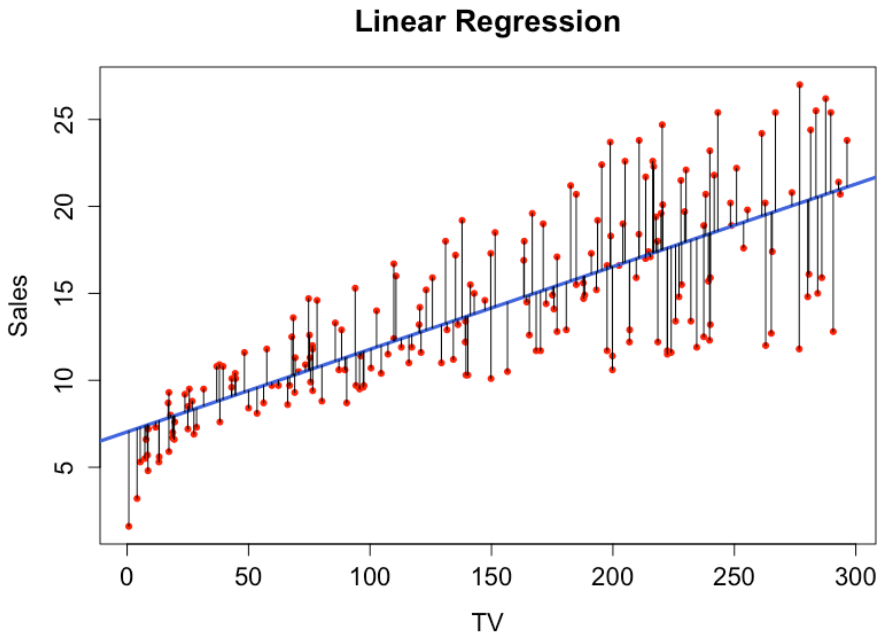

Figure 1: This plot depicts minimizing the least squares criterion in simple linear regression. Each red dot in the plot corresponds to a data point in our training set. The blue line is the predicted output line obtained from our simple linear regression model. The vertical lines between the red dots and the blue line are the errors our prediction makes. These are the very same errors that we are minimizing via least squares (specifically, the it is the squares of these errors that we are minimizing). Thus, the observed blue line in the plot is the line that best minimizies these squared errors: it is the line that fits the closest to the data points in our training set.

$$H_0 : There is no relationship between X and Y$$

$$H_1 : There is some relationship between X and Y$$

Mathematically, we can construct our tests in the following form:

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta 1! = 0$$

Intuitively, if the value of *beta_1* is 0, that means changing our value of X will have no effect on our obtained outcome. On the other hand, if that value is not equal to 0, then we know that changing X will result in some change in Y, in which case some relationship (no matter the magnitude) does exist.

Because we're performing a simple linear regression and only have two variables, a t-test will suffice to measure the relationship between X and A_Y_. We'll construct our t-test in the following manner:

$$t = \hat{\beta}_1 - 0/SE(\beta_1)$$

This test measures the number of standard deviations that our estimate for *beta_1* is away from 0. We measure this under the null hypothesis, dictating that t will follow a t-distrbition with n-2 degrees of freedom. In this way, we obtain our p-value, the probability of observing what we saw *by chance*, assuming the null hypothesis. P-values reveal information in the following manner:

- *Small p-value*: There is some association between the dependent and independent variables.
- *Large p-value*: There is no association between the dependent and independent variables.

Assessing Model Accuracy

After performing our regression, we can assess how accurate our model performed. In other words, how well did our model fit the data?

For a simple linear regression, we'll employ two methods: the RSE and R^2 (R-squared).

RSE

RSE is the residual standard error. It is an estimate of the standard deviation of the error in our model. Intuitively, it is the average amount that our response will deviate from the true regression line, defined as follows:

$$sqrt(1/(n-2) * \sum_{i=1}^{n} (y_i - yhat_i)^2)$$

We consider this a measure of the *lack of fit* of our model.

R Squared

Whereas the RSE measures a lack of fit, R Squared measures the goodness of fit of our model. Intuitively, it measures the proportion of variance explained by our model. Thus it has the range [0,1] and is independent of Y's scale.

$$R^2 = (TSS - RSS)/TSS$$

TSS measures the total sum of squares; the total variance inherent in the response before performing our regression. *RSS* meausres the amount of variability explained by our regression. Thus the above formula measures the proportion of variability in Y that is explained using X.

- R Squared near 1: A large proportion of variability in the response explained by the regression; i.e. our model fits the data well.
- R Squared near 0: The regression did not explain much variability in the response; i.e. our regressin performed poorly.

Although R Squared has the advantage of being more interpretable than RSE (due to it's restriction between 0 and 1), it is still problem dependent.

One last interesting property of simple linear regression is that R Squared is equal to Cor(X,Y).

Results

As stated, from the *Advertising* data set we obtain the following data set:

$$(TV_1, Sales_1), (TV_2, Sales_2), ..., (TV_n, Sales_n)$$

Thus, our simple linear regression will take the following form:

$$sales = \beta_0 + \beta_1 * TV$$

The plot obtained for our data was already shown. So how did we do?

Our Coefficient Test

```
{r echo = F} suppressMessages(library(xtable)) reg <- lm(Sales ~ TV,
data=adv) reg_table <- xtable(reg) print(reg_table, type = "latex",
file = "reg-table.tex")
```

Thus, for our data, we can see that our coefficient value for **TV** is 0.0475, In other words, increasing the **TV** advertising budget by $1,000 is associated with an increase in sales with about 48 units. Because our p-value is so small ($<$ 0.0001), we reject the null hypothesis and conclude that there is indeed some relationship between X and Y.

Our Model Accuracy

Insert table

As the table displays, our value for the RSE is 3.26, revealing that actual sales in each market deviate from the true regression line by approximately 3,260 units on average. This sounds like a lot, but it's really problem dependent. For us, the mean value of *sales* is roughly 14,000 units, so the percentage error of our regression line is 3,260/14,000 = 23%.

Our measure for R Squared is .61. So 61% of the variability of *sales* is explained by a linear regression on TV.

Thus, while our simple linear regression model wasn't perfect, it wasn't horrible. To achieve better predictive accuracy, a more complicated model should be employed. That said, a simple linear regression is a good starting point for any predictive task that assumes some linear structure underpinning the data.

Ask Stewart: - How to add est hat - Why plot shows below when I convert