Simple Regression Analysis

Timothy Park

Oct 07,2016

Abstract

Having learned the basic tools used in workflow reproducibility, the next step is to start applying this computational toolkit and R programming to reproduce a simple regression analysis. This report will replicate the regression analysis performed in Chapter 3 (Linear Regression) of the textbook, "An Introduction to Statistical Learning".

Introduction

This analysis involves fitting a linear model to the Advertising data set, regressing Sales onto TV. The regression line produced by the coefficient estimates is an effective predictor for test sets, assuming the observed data is approximately linear. The accuracy of the linear model will be assessed by 3 useful quantities, providing more information about predictions and inferences among the variables.

Data

The Advertising data set consists of Sales (in thousands of units) in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper. In this analysis, we will only look into the relationship between product Sales and TV advertising budgets, response and predictor variables respectively.

Methodology

We use the following linear model to evaluate the association between Sales and TV advertising: Sales = $\beta_0 + \beta_1$TV

Using the least squares criterion, estimates of the beta coefficients are calculated to form a fitted regression line from the given n = 200 observations.

Results

Visualizations displayed below are the results of the simple regression model fitted onto the Advertising data set.

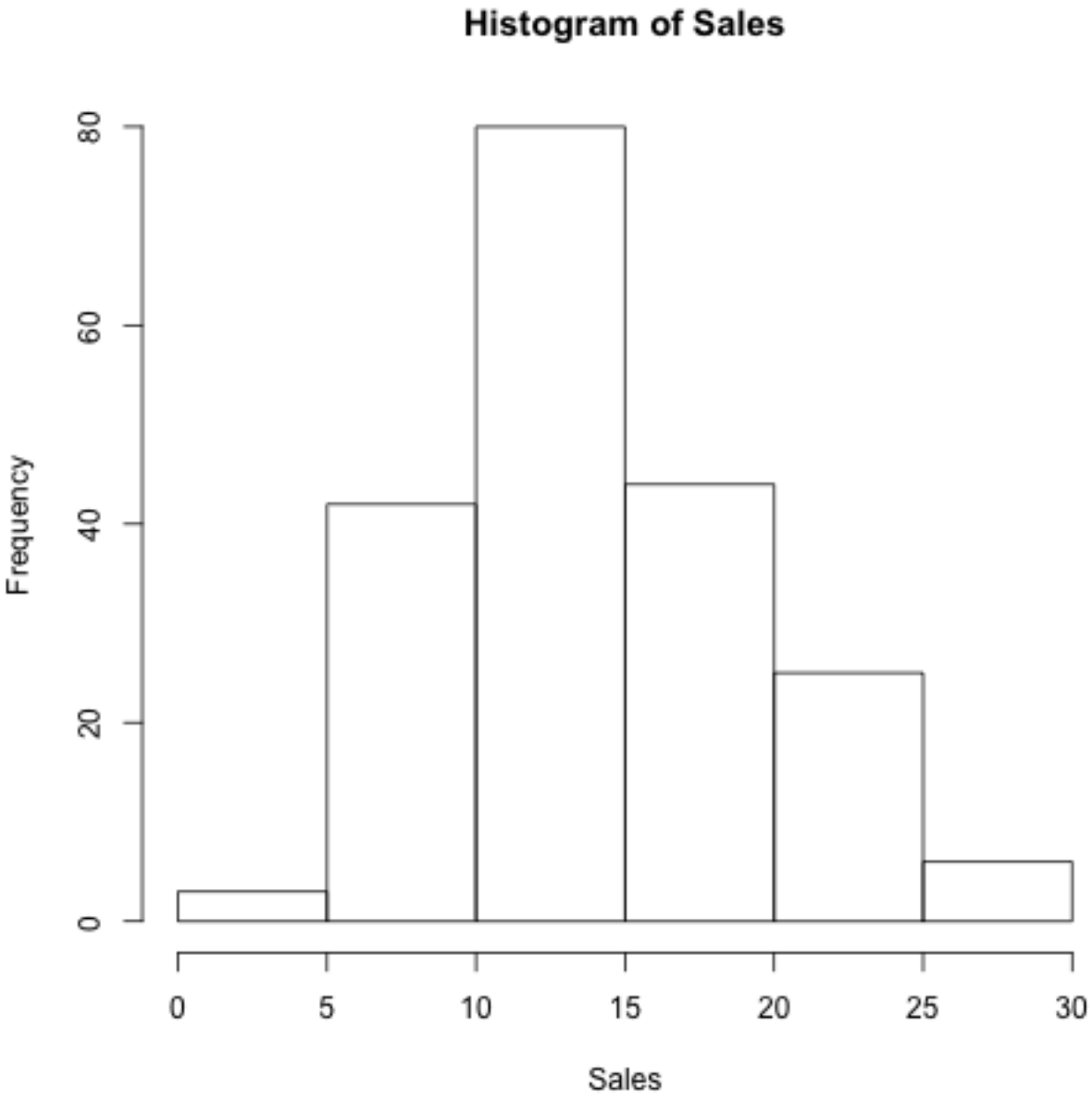

Figure 1: Histogram of Sales

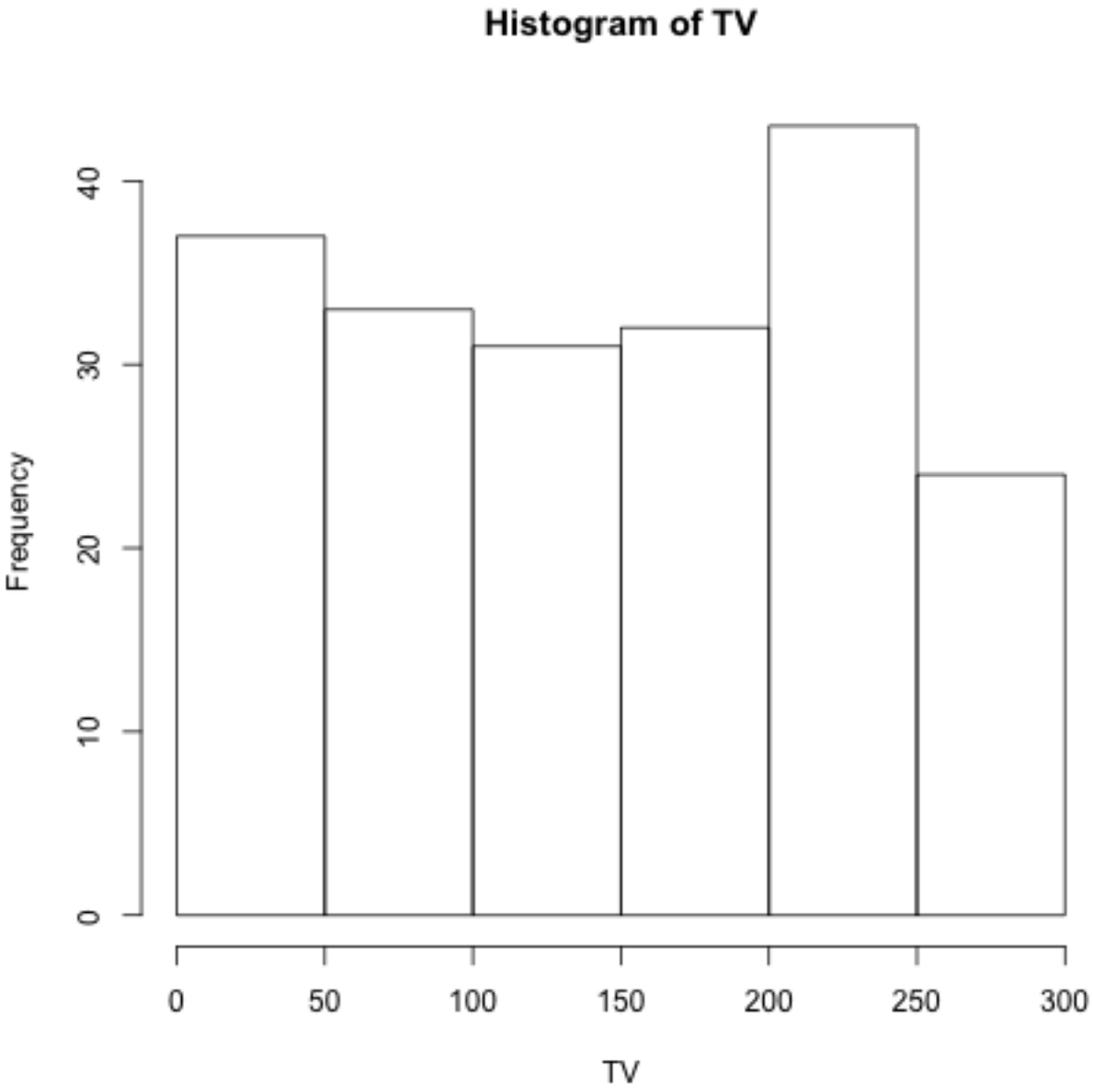

Figure 2: Histogram of TV Advertising

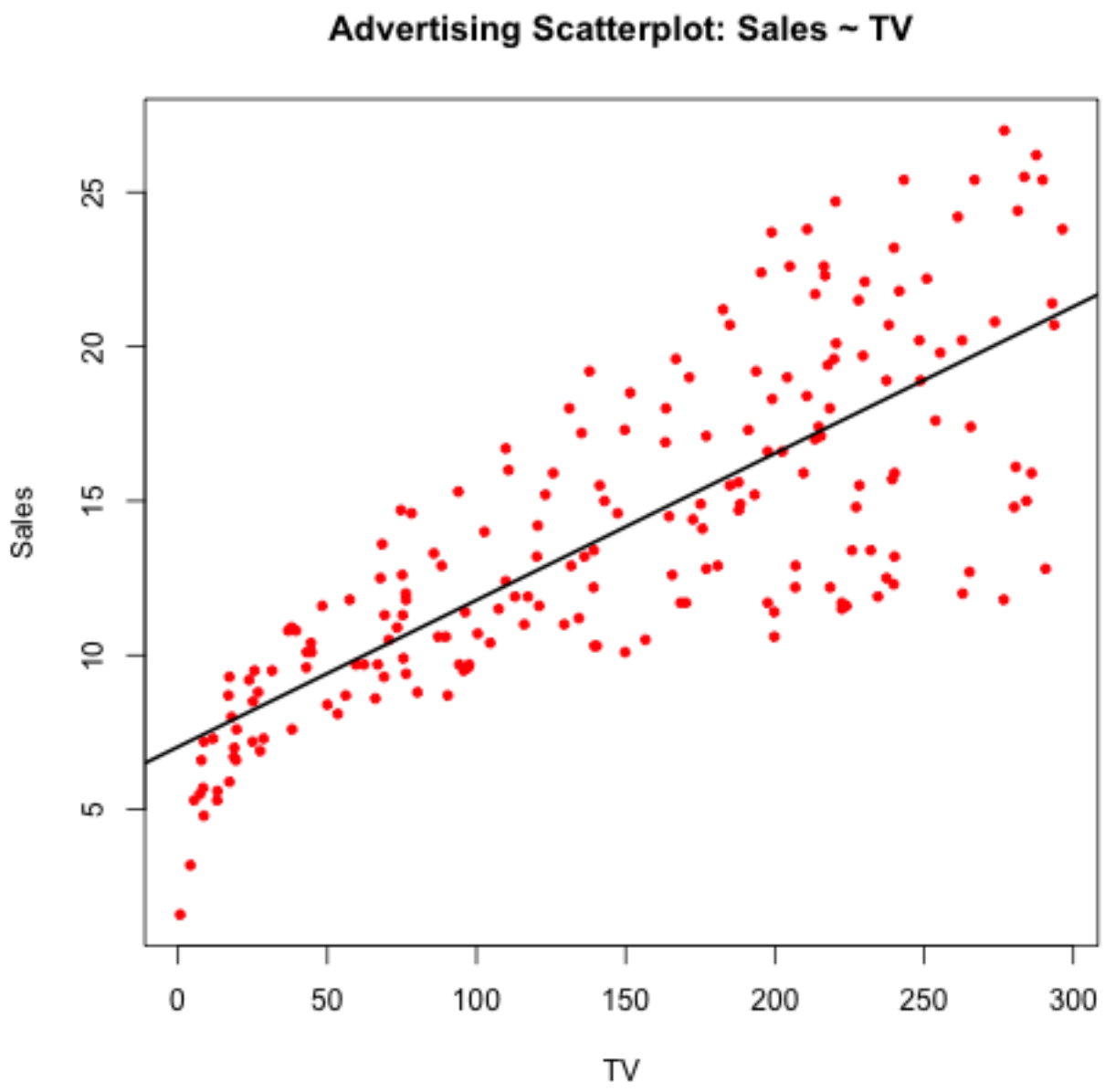

Figure 3: Scatterplot of all Advertising observations and the fitted linear regression line

Fitting the model using the `lm()` function, a summary of the beta estimates and other statistics are shown below.

Table 1: Information about Regression Coefficients

Coefficients	Estimate	Std. Error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	<0.0001
TV	0.0475	0.0027	17.67	<0.0001

The table below displays the quantities to assess the accuracy of the linear model fit to the training data.

Table 2: Regression Quality Indices

Quantity	Value
Residual Standard Error	3.259
R-squared	0.612
F-statistic	312.1

Conclusions

The histograms of the TV and Sales variable does not provide much information about their association; however, a linear fit tells otherwise. Evaluating the variables' distribution, Sales and TV looks approximately normal and uniform, respectively.

The scatterplot displays the Sales observations in 200 different markets in relation to TV advertising. From the data points alone, it is safe to assume the relationship to be linear, so a linear model fit is appropriate. A linear regression line is superimposed onto the scatterplot and the beta coefficients are generated via the least squares criterion. This linear fit seems to represent the relationship between the response and the predictor, but the variance of Sales variable might produce a high residual standard error and thus a low R-squared statistic.

In table 1, both standard errors of each beta coefficient is small in comparison to their respective estimates. These relationships imply a large t-statistic and thus a low p-value. A p-value less than 0.0001 indicates that it is justified to reject the null hypotheses that each beta coefficient is equal to 0. This conclusion about β_1 means that TV advertising has a strong association to the quantity of Sales. More specifically, a \$1000 increase in TV advertising corresponds to a 47.5 additional increase in Sales. For β_0, assuming that the budget ignores TV advertising (X = 0), the quantity of Sales will be equal to 7032 units by default.

In table 2, the residual standard error (RSE) is the mean of the residual sum of squared errors for every observation. The lower RSE, the stronger the fit. However, if the RSE is too low, then the model might be a case of overfitting, so using this specific model for test data may lead to high RSEs. An RSE of 3.259 means that the linear model will, on average, have a prediction error of 3259 in Sales.

The R-squared quantity is inversely related to the RSE; a low RSE coincides with a high R-squared statistic. A R-squared of 0.612 does not infer a strong fitting of the model, so a more flexible statisical learning method is recommended.

The F-statistic tests the null hypothesis that all the beta coefficients have no association to the response variable Sales. A high F-statistic of 312.1 rejects the null hypothesis and suggests that at least one of the betas have an association to Sales, which is evident in the first table displaying the low p-values of the beta estimates.

Based on the assessment quantities of RSE and R-squared, it is apt to suggest a different model to fit the Advertising data set. However, though not a proper fit, there is solid evidence of a strong association between TV and Sales. A low p-value for both the F-statistic and TV beta estimate confirms this statement.