Simple Regression Analysis

Morgan Smart

10/9/2016

Abstract

This paper replicates the analysis from Section 3.2 of Chapter 3. *Linear Regression* in "An Introduction to Statistical Learning" by James *et al.* Section 3.2 looks at advertising data and assesses: if at least one of the predictors useful in predicting the response, if all predictors help to explain the response or if only a subset of the predictors are useful, how well the model fits the data, and how accurate the model prediction is. In Section 3.2 and in this paper, these questions are answered by computing a multiple linear regression of TV, Radio, and Newspaper advertising budgets (in thousands) on Sales (in thousands) and analyzing the regression results.

Introduction

A multiple linear regression is an approach to predicting a quantitative response Y based on a multiple predictor variables X_1 through X_p, where Y and X_1 through X_p are vectors and each value in each $X_(ij)$ ($x_(ij)$) has a corresponding value in Y (y_i). The model assumes that the relationship between every X_i and Y is linear; thus, in order to compute a linear regression that has an accurate interpretation, every X_i and Y **must** have a linear relationship. Additionally, the model assumes that each X_j isn't correlated with any other X_j. The multilple linear model can be written as $Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon$, where β_0 is the intercept and β_1 through β_p are the slopes of their corresponding predictor variable X_j. These beta values are all constants, unknowns, and together are the model coefficients. The interpretation of β_0 is the expected mean value of Y without a predictor variable and the interpretation of β_1 through β_p is the change in Y for a unit increase in the beta's corresponding X_j. Although betas are unknown, we can estimate them using the multiple linear regression model: solving for the intercept and slopes that produce the plane closest to each point ($x_(ij)$,y_i) in each X_j,Y. Once we have an estimate for the betas, have verified that none of the X_j are correlated, and have verified that the relationship between each X_j and Y in linear, we can compute a multiple linear regression to determine the strength of the relationship between each X_j and Y, if the relationship is statistically significant, and how accurately the model predicts the relationship. This process will be illustrated in the following sections using the Advertising dataset presented in Section 3.2 of Chapter 3. *Linear Regression* in "An Introduction to Statistical Learning."

Data

The Advertising dataset has $n = 200$ row entries (data points) and 5 columns. These columns are:

- X = the row index
- TV = Advertising budget for TV (in thousands)
- Radio = Advertising budget for Radio (in thousands)
- Movies = Advertising budget for Movies (in thousands)
- Sales = Product sales (in thousands) made having the respective TV, Radio, and Movie advertising budget

The row values for each column (other than column X) are in thousands so interpretation of the multiple linear model of TV, Radio, and Newspaper advertising budget on Sales is more straight forward. For the purposes of replicating the analysis done in Section 3.2, we will remove the row index column "X" because it is not used in the multiple linear regression. To understand the data, below are histograms of each predictor variable (TV, Radio, and Newspaper) and also of the dependent variable (Sales).

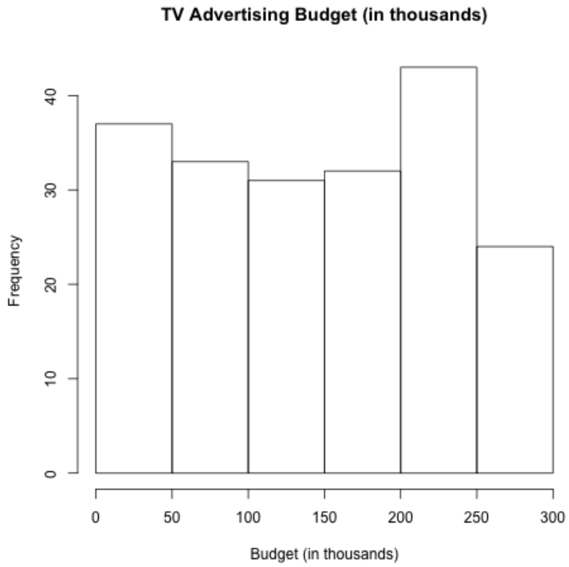


Figure 1: TV Ad Budget Histogram

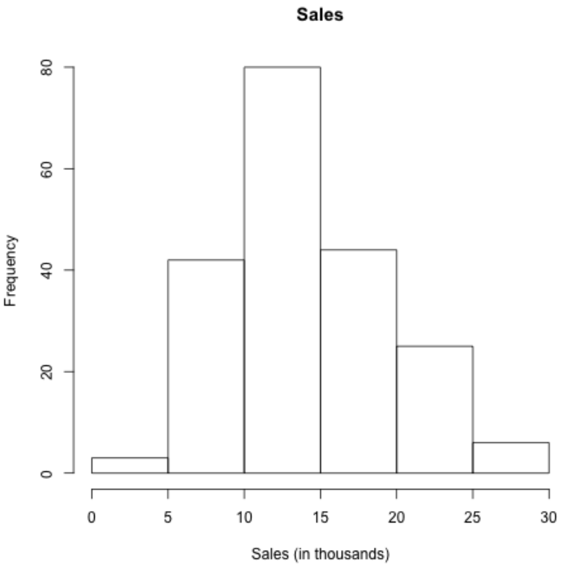


Figure 2: Product Sales Histogram

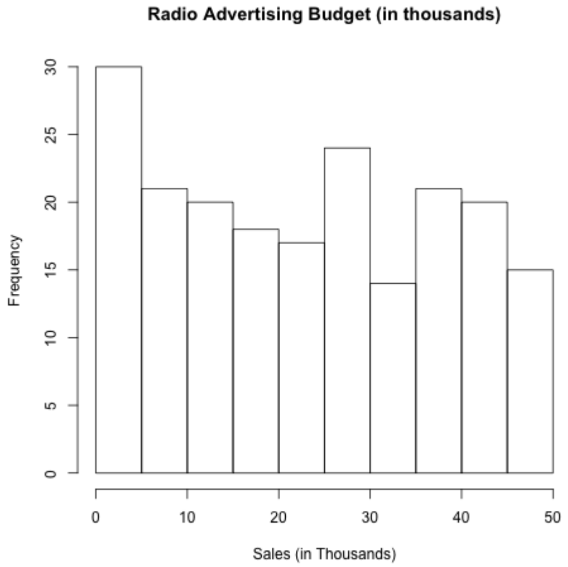


Figure 3: Radio Ad Budget Histogram

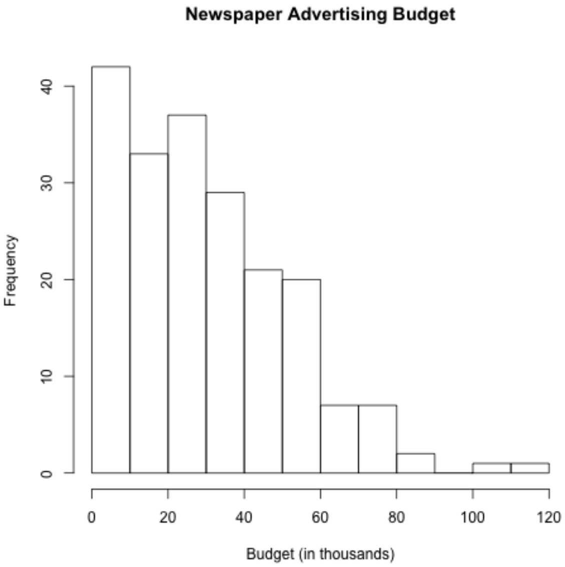
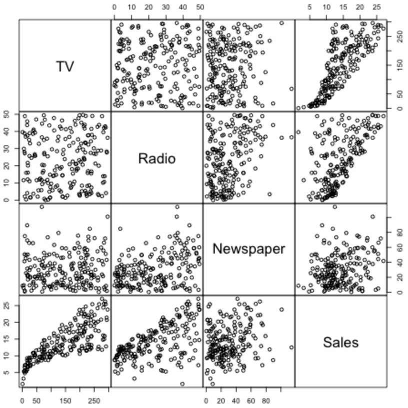


Figure 4: Newspaper Sales Histogram

As evident from Figures 1, 3, and 4, the largest range of ad budget in the data set is for TV, second largest is for newspaper, and smallest is for radio. It's important to keep these ranges in mind when using the multiple linear regression model so that we don't extrapolate a prediction beyond the values in our data set. We also see from Figure 2 that the dependent variable (Sales) is normally distributed, an assumption of the multiple linear regression. Another assumption of the model is that the dependent variable has a linear relationship with each predictor variable. To test this, we plot a pairwise scatterplot below in Figure 5.

As evident from Figure 5, each predictor variable (TV, Radio, and Newspaper) shares a linear relationship with the dependent variable (Sales). We also must verify that the predictor variables are independent with

Figure 5: Pairwise Scatterplot of all Variables



one another. This is somewhat clear from Figure 5 given there is no clear relationship between any of the predictors in the plots, but to further verify that the predictors are in fact independent, we calculate their correlation matrix (Table 1 below).

Table 1:

	TV	Radio	Newspaper	Sales
TV	1	0.055	0.057	0.782
Radio	0.055	1	0.354	0.576
Newspaper	0.057	0.354	1	0.228
Sales	0.782	0.576	0.228	1

From Table 1, we see the highest correlation between any of the predictor variables (TV, Radio, Newspaper) is that of Newspaper and Radio (0.354) which is low. Thus, the predictor variables aren't too correlated and we can use the multiple linear regression to assess the effect of the predictors on the dependent variable (Sales).

Methodology

To assess if there is a relationship between TV, radio, or newspaper advertising budget and product sales, we ran both a multiple linear regression of all of the ad budget types on product sales as well as a simple linear regression of each ad budget type on product sales–as computed in Section 3.2. We then assessed the statistical significance of the relationship between each of the ad budgets on and product sales using the p-value of the predictor in the simple linear regression case and also when among the other predictors in the multiple linear regression case. If the p-value of the predictor variable X_j is less than 0.05, X_j has a statistically significant effect on the response variable Y. Finally we looked at the multiple regression summary plots as well as the Residual Standard Error (RSE), the R^2, and the $F-statistic$ of the multiple regression model to asses the models accuracy. Because we estimate the betas in our model, every observation has an error term (ϵ) associated with it. The RSE measures the standard deviation of ϵ: the average amount the predicted response will deviate from the true regression line. It's hard to interpret what a good RSE is because RSE is measured only in the units of Y. Thus, we also look at R^2 which measures the proportion of variance explained by the data (always a value between 0 and 1). Ideally, we'd like to see an R^2 that is close to 1, though this may not always be realistic depending on the question being asked. We can also assess our model's accuracy by looking at the $F-statistic$ which tests the full model (including all of the predictor variables X_j) against the minimalist model that assumes all values of each X_j to be zero and uses only the mean of the values in Y (β_0). Associated with an $F - statistic$ is a p-value. If this p-value is less than 0.05, there is little chance that the values of the predictor variables X_j are zero and thus the full model is more accurate than the minimalist model.

Results

The simple linear regression of TV advertising budget on product sales tells us that TV advertising budget has a statistically significant effect on product sales since the p-value of TV (0) is less than 0.05. This p-value can be seen in the summary statistics of the simple linear regression in Table 2. From Table 2, we also see that a $1,000 increase in TV advertising budget increases product sales by 4.8%.

Table 2:

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

Table 3:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.312	0.563	16.542	0
Radio	0.202	0.020	9.921	0

Table 4:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.351	0.621	19.876	0
Newspaper	0.055	0.017	3.300	0.001

The simple linear regression of Radio advertising budget on product sales tells us that Radio advertising budget also has a statistically significant effect on product sales since the p-value of Radio (0) is less than 0.05 (seen in Table 3). Table 3 also shows that a $1,000 increase in Radio advertising budget increases product sales by 20.2%. The simple linear regression of Newspaper advertising budget on product sales tells us that Newspaper advertising budget also has a statistically significant effect on product sales since the p-value of Newspaper (0) is less than 0.05 (see Table 4). Also from Table 4, we see that a $1,000 increase in Radio advertising budget increases product sales by 5.5%.

Table 5:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.939	0.312	9.422	0
TV	0.046	0.001	32.809	0
Radio	0.189	0.009	21.893	0
Newspaper	-0.001	0.006	-0.177	0.860

Although each ad budget has an individual significant effect on product sales respectively, we'd like to assess the effect (if any) the ad budgets have on product sales when they all are taken into account. To do this, we ran a multiple linear regression of TV, Radio, and Newspaper ad budget on product sales. The results can be seen in Table 5 above. From Table 5, we see that both Radio and TV ad budget still have significant effect on product sales since their respective p-values (0 and 0) are less than 0.05. Newspaper ad budget, however, is no longer significant when assessed among TV and Radio budget, having a p-value now greater than 0.05 (0.86)

Finally, we assessed the accuracy of the multiple regression model and are confident that it fits the data well based on its RSE, R^2, and $F - statistic$ (Table 6) and by looking at the Normal QQ Plot, Residual Plot, and Scale Location Plot of the model (Figures 6, 7, and 8).

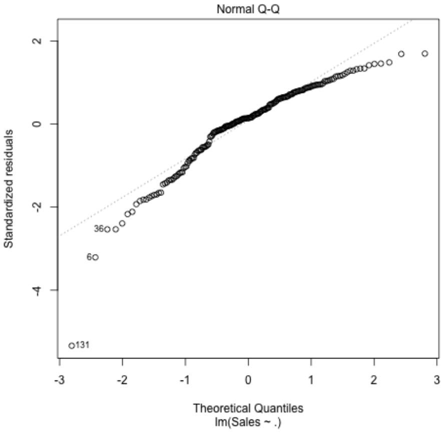


Figure 6: Normal QQ Plot

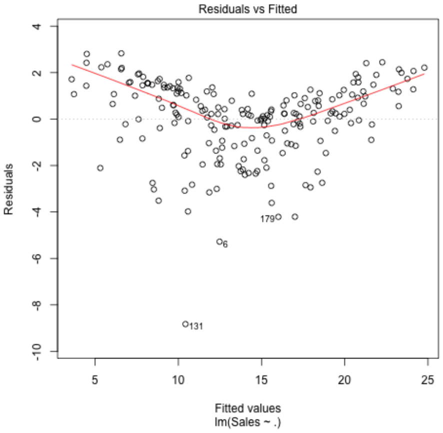


Figure 7: Residual Plot

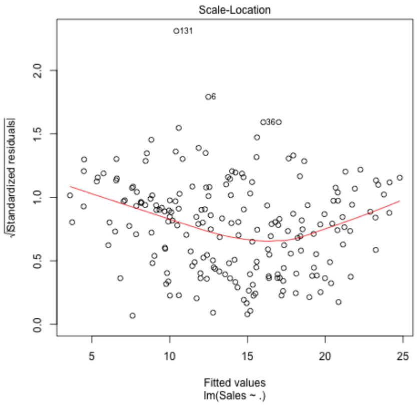


Figure 8: Scale Location Plot

Looking at the Normal QQ Plot, we see the tails both dip down below the line. In a perfect world, we'd like to see all the point fall on the line. This would mean the data are exactly normal. However, the data are close to the line so we move on to inspect other elements of fit as well. From the Residual Plot, it seems the residuals are not homoscedastic (ideal for the model) since they don't fall evenly around and along the line. However, when we look at the Scale Location Plot, the residuals become more evenly distributed and thus we believe the model is a good fit in this sense.

We also look at the summary statistics of the model to get a better idea of it's fit. We see that the RSE is 1.686, meaning that the prediction of product sales based on TV advertising budget is off by 1686 units on average. The R^2 of the model is 0.897 meaning roughly two-thirds of the variability in product sales is

Table 6:

	Quantity	value
1	Residual standard error	1.686
2	R squared	0.897
3	F-statistic	570.271

explained by the model. Finally, the $F - statistic$ is 570.3 and it's associated p-value is 0, meaning there is very little chance that the minimalist model is moe accurate than the full multiple linear regression model.

Conclusions

From this analysis, we learned that the analysis in Section 3.2 of Chapter 3. *Linear Regression* in "An Introduction to Statistical Learning" is reproducible–since we produced its results in this paper. Additionally, by viewing the data and methodology behind the analysis in Section 3.2, we verified that the assumptions of a multiple linear regression were met by the data, and so, the analysis has accurate interpretation. We also learned that a simple linear model may show a predictor as being statistically significant, but when using a more complex model, that predictor may no longer be significant. Thus, it's important to experiment with adding multiple predictor variables (when applicable, makes sense, and does not violate the model's assumptions) in order to get the most accurate results.