HW 3: Multiple Regression Analysis

Todd Vogel

10/14/2016

Abstract

The following reports looks into data on advertising costs across multiple media including television, radio, and newspaper as well as sales. By utilizing linear modelling we effectively establish the relationship between advertising and sales numbers, specifically focusing on television, newspapers, and radio. We conclude based on regression metrics such as R^2, F-Statistic, and Residual Standard Error that the model is not and effective predictor of future sales. The results are a reproduction of the work in Chapter 3.2 in the book **An Introduction to Statistical Learning.**

Introduction

Data

The data utilized in the project is exclusively an Advertising data set with 4 variables: `TV`, `Radio`, `Newspaper`, and `Sales`. `Sales`, our dependent variable is measured in thousands of units and `TV`, `Radio`, and `Newspaper` are our independent variables measured in thousands of dollars. Each data point, or row, represents an individual product (of which there are 200, in as many different markets).

Methodology

In multiple regression the goal is to look at a certain number of parameters (independent variables) and determined their effect on the dependent variable through a coefficient β. To do so we apply the general multiple regression equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$$

The completion of this project required many steps and methodologies. However, the most important is the process of regression, mentioned above. To determine the relationship between television, radio, newspaper, and sales we applied the linear model, determined from multiple regression:

$$Sales = \beta_0 + \beta_1 TV + \beta_2 Radio + \beta_3 Newspaper$$

Then, using the lm function in R, we calculate the values for the coefficients β_0, β_1, β_2, and β_3 (through the Ordinary Least Squares calculation). The remainder of the project was constructed using command line and makefile in order to ensure reproducibility. The file and directory structure was created in the terminal and r scripts were intertwined to create a reproducible output with Makefile and the `make` function.

Results

After applying the multiple regression model above certain values are calculate, the results of which are shown below.

Table 1: Information about Regression Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.94	0.31	9.42	0.00
TV	0.05	0.00	32.81	0.00
Newspaper	-0.00	0.01	-0.18	0.86
Radio	0.19	0.01	21.89	0.00

This table presents the basic regression fit between the independent variables (`TV`, `Newspaper`, and `Radio`) and the dependent variable (`Sales`) determined from the `lm()` function. β_1, β_2, and β_3 were calculated to be 0.05, -0.00, and 0.19 (for `TV`, `Newspaper`, and `Radio` respectively)

Table 2: Regression Quality Indices

Value	Quantity
RSE	1.69
R2	0.90
F-Stat	570.27

This table presents regression quality index values. By running the function `summary(lm())` we gather values for the RSE, R^2, and F-Statistic (1.69, 0.90, and 570.27 respectively).

Table 3: Correlations between Variables

	TV	Radio	Newspaper
TV	1.00	0.05	0.06
Radio	0.05	1.00	0.35
Newspaper	0.06	0.35	1.00

This final table represent the correlations between the independent variables. These numbers range from 0 to 1, 1 being perfectly correlated and 0 being uncorrelated. It is clear by looking at these values that these variables are not particularly correlated to one another as the highest value is .35 between `Radio` and `Newspaper.` This means that if any particular variable proves to be an ineffective predictor of `Sales` numbers, that does not skew, to a great extent, the predictive power of other variables.

Conclusions

In order to draw effective conclusions I must answer the 4 following questions:

1. Is at least one of the predictors useful in predicting the response?
2. Do all predictors help to explain the response, or is only a subset of the predictors useful?
3. How well does the model fit the data?
4. How accurate is the prediction?

Question 1

In order to determine if at least one of the predictors is effective we must look at the F-statistic.. Previously I concluded that the F-statistic value was approximately 570. Now, an F value of 1, or close to 1, would mean that the null hypothesis (that no predictors are significant) is true. However, because the number 570 >> 1 suggests that the null hypothesis is untrue. Therefore it is safe to conclude that at least one of the variables is a significant predictor of `Sales`.

Question 2

Now it is important to determine if a subset of predictors is useful. In this case, it could be true that all predictors are associated with the response. In fact, the more variables in the model the better the fit. However, in this case we look to the p-values to determine if all of the variables are significant. Because the p-value for `Newspaper` is not significant ($p = 0.86$), it is fair to conclude that a subset of predictors are useful (`TV` and `Radio`).

Question 3

In answering whether the model fits the data well, we must look to the R^2. The closer an R^2 value is to 1, the more a model explains the response variation. Therefore, the closer R^2 is to 1, the better the fit of the model. Becuase we calculated R^2 to be .8972 with our model, it is clear that the model fits the data well.

Question 4

In the end the prediction is not that accurate as the prediction interval is very large. This is because of model bias and irreducible error. Large standard errors in our calculation lead to innaccurate predictions.