Multiple Regression Analysis

Ellen Hwang

10/12/2016

Abstract

This report will be examining the applications of multiple regression analysis. We will also be looking into more methods of finding relationships among variables. I am using the Advertising dataset, taken from the webpage for the text, *An Introduction to Statistical Learning*, to perform this simple linear regression. This report will include a description of the data, methodology, and results for the multiple linear regression.

Introduction

Multiple Linear Regression allows one to predict the relationship between one dependent variable and several independent variables. In this report, I was particularly examine the relationship the effect of *TV*, *Newspaper*, and *Radio* budgets on *Sales* revenues. I will discuss my methods in R, my various results, and conclusions about the relationships and statistics I find.

Data

We will be working with the Advertising dataset. This dataset holds information for 200 different markes for 4 different variables: *Sales*, *TV*, *Newspaper*, and *Radio*. *Sales* represents the amount of units sold (in thousands). *TV*, *Newspaper*, and *Radio* each represent the advertising budget spent on those platforms.

Methodology

For running a regression, we using the `lm()` function. More specially, we regress *Sales* on *TV*, *Newspaper*, and *Radio* using this code: `lm(Sales ~ TV + Newspaper + Radio, data = advertising)`. The immediate output of this code is the estimated values on each coefficient and the constant value. Second, I use the `summary` function on the regression object to see further information about the coefficients , residuals, and other statistics.

Results

	TV	Radio	Newspaper	Sales
TV	1.00	0.05	0.06	0.78
Radio	0.05	1.00	0.35	0.58
Newspaper	0.06	0.35	1.00	0.23
Sales	0.78	0.58	0.23	1.00

Table 1: Correlation Matrix for all Variables

Prior to starting the analysis, it is good practice to examine the correlations among all variables. As part of the analysis, I will also answer a couple 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 3 predictors useful?
3. How well does the model fit the data?
4. How accurate is the prediction?

1. Is at least one of the predictors useful in predicting the response?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.0326	0.4578	15.36	0.0000
TV	0.0475	0.0027	17.67	0.0000

Table 2: Regressing Sales on TV

Based on the regression of Sales on TV, we found that TV is a statistically significant variable because it has a p-value of 0 which is very close to zero. The R squared is 0.61 meaning about 61.19 % of the variance from the population regression line is explained by Radio. We can also see that the coefficient on TV is a postive value, meaning TV and Sales are positively correlated.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.3514	0.6214	19.88	0.0000
Newspaper	0.0547	0.0166	3.30	0.0011

Table 3: Regressing Sales on Newspaper

Based on the regression of Sales on Newspaper, we found that Newspaper is a statistically significant variable because it has a p-value of 0.0011 which is very close to zero. The R squared is 0.05 meaning about 5.21 % of the variance from the population regression line is explained by Radio. We can also see that the coefficient on Newspaper is a postive value, meaning Newspaper and Sales are positively correlated.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.3116	0.5629	16.54	0.0000
Radio	0.2025	0.0204	9.92	0.0000

Table 4: Regressing Sales on Radio

Based on the regression of Sales on Radio, we found that Radio is not a statistically significant variable because it has a p-value of 0 which is very close to zero. The R squared is 0.33 meaning about 33.2 % of the variance from the population regression line is explained by Radio. We can also see that the coefficient on Radio is a postive value, meaning Radio and Sales are positively correlated.

Overall, we see that based on the individual regressions, at least one variable is useful in predicting the response.

2. Do all predictors help to explain the response, or is only a subset of the predictors useful?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.9389	0.3119	9.42	0.0000
TV	0.0458	0.0014	32.81	0.0000
Newspaper	-0.0010	0.0059	-0.18	0.8599
Radio	0.1885	0.0086	21.89	0.0000

Table 5: Regressing Sales on TV, Newspaper, and Radio

When we run a regression on more than one variable, we find that not all predictors. Both TV and Radio have a p-value of less than .05 meaning they are statistically significant explanatory variables. On the other hand, Newspaper has a p-value of 0.86. This value is above the 0.05 threshold, meaning it is not a statistically significant explanatory variable. We can then conclude that only a subset of the predictors, TV and Radio, help explain the response.

3. How well does the model fit the data?

	Quantity	Value
1	Residual standard error	1.69
2	R squared	0.90
3	F-statistic	570.27

Table 6: Strength of Relationships Statistics

We can look to the R squared statistic and the residual standard error to examine this question. Our R squared is at about 0.9, meaning the variables explain about 90% of the variance in Sales. The RSE is at about 1.69 meaning by dividing RSE by the Sales mean, we get the percentage error of 12.02%.

4. How accurate is the prediction?

In this case, we would have to split the data into training, validation, and testing datasets to answer this question. Using various values for the explanatary variables, Radio, Newspaper, and TV, we would create several confidence intervals or prediction intervals. Subsequently, we would account for the number of predictions within these CIs or PIs to get our accuracy of the model.

Conclusion

From running multiple regression, we can observe the significance of each variable on the dependent variable. In our report, we found that although the F-statistic told us at least one of the explanatory variables would explain Sales, not all are statistically significant to the regression. Moving forward, we would need to remove Newspaper from the regression to get a better model R squared and RSE. This report has shown the ways we can answer important questions about data using multiple linear regression.