International Journal of Business and Management Invention (IJBMI)
ISSN (Online): 2319 – 8028, ISSN (Print): 2319 – 801X
www.ijbmi.org || Volume 7 Issue 6 Ver. II || Jun. 2018 || PP—73-77

A Machine Learning Analysis Of Advertising In Magazine

Andy W. Chen[1]
[1]*Sauder School of Business, University of British Columbia, Canada*
Correspondence Author: Andy W. Chen

ABSTRACT: *In this paper, I investigate the factors that cause magazine advertising costs to vary using machine learning models. In particular, I build ordinary least squares (OLS) regression models to predict magazine advertising costs given features such as magazine circulation, income of readers, and gender. I evaluate the models using the paired t-test and adjusted R^2. The adjusted R^2 is a more informative measure than the commonly used R^2. I find circulation and median income of readers to be significant. The adjusted R^2 is found to be 0.6734, suggesting that 67.34% of the variation in advertising costs can be explained by the independent variables.*

KEYWORDS: *Advertising, Business Strategy, Machine Learning, Linear Regression*

Date of Submission: 11-06-2018 Date of acceptance: 26-06-2018

I. INTRODUCTION

The price of advertising varies from one magazine to another. What causes the prices to vary? There may be multiple reasons. Magazines that reach more readers should reasonably charge more for advertising. Also, magazines with readers having higher income should theoretically be able to charge more. In this paper, I propose a machine learning model to provide some answers that can help explain the impact of circulation, income and gender on advertising prices. I build an ordinary least squares (OLS) regression model, which is a machine learning approach for training a model that can be used to predict the target variable (advertising price) given the feature variables (circulation, income and gender).

I then evaluate the significance of each feature using the paired t-test and adjusted R^2. The t-tests show that that the feature variables circulation and median income are important in the model. The adjusted R^2 is 0.6734, meaning that 67.34% of the total variation in the dependent variable (page cost) is explained by all of the independent variables (circulation, percentage of male readership, and median income) taken together, after adjusting for the number of independent (predictor) variables.

Past research in this area includes the work by James and Vanden Bergh[1], who find that direct response advertisements would differ significantly from institutional advertisements in the amount of information and type of information. Healey and Kassarjian[2] find that the number of attribute claims in magazines decreased following the FTC (Federal Trade Commission) program. Fuchs[3] conduct an experiment and found significant impact of magazine and sponsor's prestige on advertised products. Soley and Reid[4] explore consumer satisfaction with information contained in magazines and found that middle-income consumers would be more satisfied with the informational value in magazines than the ones with low and high income. Morris et al.[5] study prescription drug advertisements and found that magazine advertisements would lead to greater patient authority in purchasing decisions of drugs.

II. METHODS

I gather a dataset of page cost, circulation of magazines, gender of readers of the magazine, and median income of readers. There are 48 magazines in total. Table 1 shows the data used.

Magazine	Page Costs (color ad)	Circulation (thousands)	Percent Male	Median Income
Better Homes & Gardens	$73,820	8,000	22	$23,241
Business Week	$35,140	845	72	$30,884
Car & Driver	$23,795	725	88	$25,982
Cosmopolitan	$28,980	2,250	17	$22,785
Ebony	$21,886	1,250	42	$16,505
Family Circle	$62,750	7,450	13	$21,785
Field & Stream	$33,760	2,000	79	$24,337
Forbes	$25,090	700	74	$36,783
Fortune	$30,040	670	71	$35,204
Glamour	$24,340	1,800	6	$21,828
Golf Digest	$26,625	1,025	82	$32,949
Good Housekeeping	$58,020	5,000	13	$21,980
Harper's Bazaar	$16,200	650	8	$25,358
Hot Rod	$20,400	850	78	$23,056
House & Garden	$25,430	1,000	19	$23,726
House Beautiful	$19,775	800	16	$24,198
Ladies' Home Journal	$48,000	5,000	12	$21,583
Mademoiselle	$16,280	1,000	9	$23,660
McCalls	$59,830	6,200	12	$20,690
Mechanix Illustrated	$24,815	1,600	81	$22,568
Money	$25,740	1,000	58	$31,587
Motor Trend	$21,905	750	84	$23,878
Ms.	$7,845	450	16	$24,107
National Enquirer	$26,500	4,637	40	$19,969
National Geographic	$95,575	8,400	55	$26,294
National Lampoon	$9,900	580	72	$22,888
Newsweek	$63,850	2,950	62	$26,719
Outdoor Life	$28,475	1,500	82	$23,596
Parents Magazine	$36,960	1,650	22	$20,779
People	$56,425	2,350	40	$23,971
Playboy	$55,710	5,000	78	$24,051
Playgirl	$7,220	650	36	$19,329
Popular Mechanics	$26,932	1,600	81	$25,474
Popular Science	$26,820	1,800	80	$26,542
Reader's Digest	$97,700	17,900	44	$21,802
Redbook	$42,675	3,800	11	$22,794
Road & Track	$18,775	630	89	$28,093
Rolling Stone	$17,770	700	66	$24,074
Scientific American	$24,000	720	68	$29,531
Seventeen	$19,250	1,500	9	$21,251
Sports Illustrated	$54,165	2,250	80	$26,275
The Star	$21,350	3,400	35	$19,156
Time	$85,870	4,400	56	$26,908
True Story	$13,435	1,400	18	$14,325
TV Guide	$77,400	17,345	45	$20,461
U.S. News and World Report	$42,510	2,050	63	$26,998
Vogue	$18,000	950	11	$23,452
Woman's Day	$60,435	7,125	8	$21,910

Table 1: Data on magazine page cost, circulation, gender and median income of readers

I train an ordinary least squares linear regression model. This is a common model used in the machine learning literature for modeling the effect of feature variables on the target variable. The model can also be used to predict the target variable, which is advertising cost per page, given values of the features. In particular, I estimate parameters in the following model.

Page Cost = β_1 Circulation + β_2 Percent Male + β_3 Median Income

The results will who the magnitude of each feature variable on advertising page cost. I evaluate the model with statistical tests such as the paired t-test, partial F-test, and adjusted R^2.

III. RESULTS AND DISCUSSION

I first present the results from some data exploration. Figure 1 shows the distribution of advertising cost per page and magazine circulation. Figure 2 shows the distribution percentage of male readership and median income of readership.

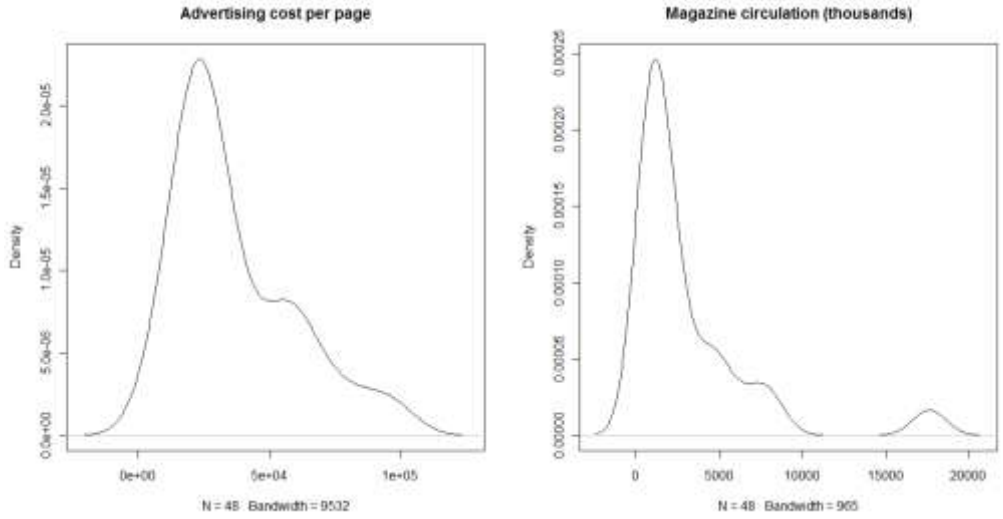

Fig. 1: Density Histograms of Advertising Cost per Page & Magazine Circulation (thousands)

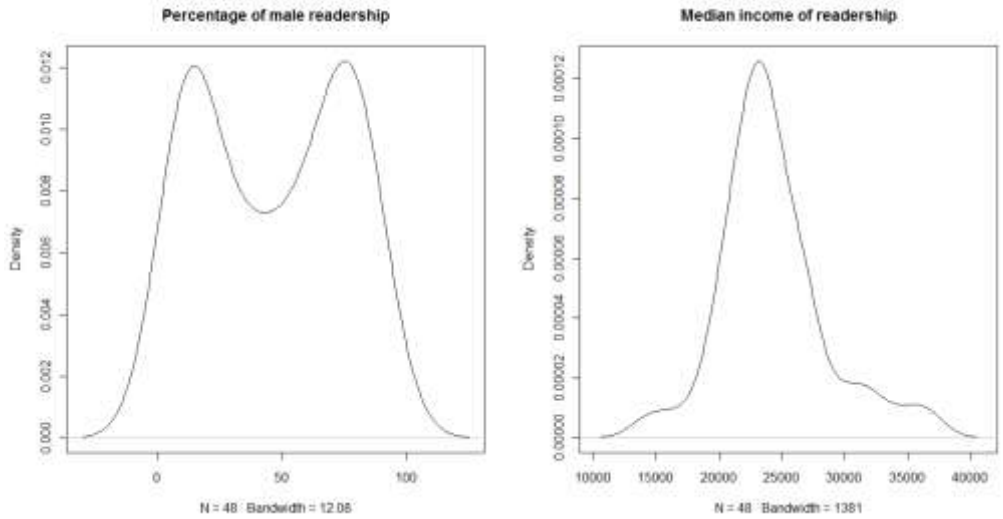

Fig. 2: Density Histogram of Percentage of Male Readership & Median Income of Readership

Scatterplots were generated in order to explore the relationships between variables. Figure 3 shows a pairwise scatterplot in order to give an initial impression of the dataset we are dealing with.

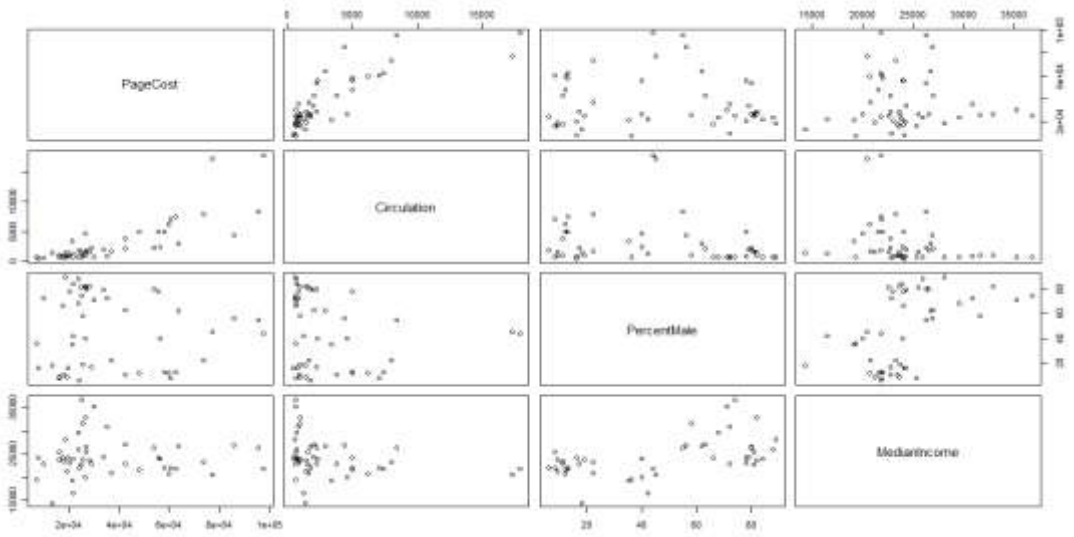

Fig 3: Pairwise Scatterplot of Page Cost, Circulation, Percentage of Male Readers, Median Income

It can be seen from these scatterplots that there exists a positive (and relatively strong) linear relationship between advertising cost per page and the level of magazine circulation, though this scatterplot does contain some outliers. We do not, however, immediately see clear linear relationships between the advertising cost and our two other predictor variables (percentage of male readership and median income). There is an initial indication towards potential collinearity between the percentage of male readership and median income. Certainly these apparent relationships give motivation for us to further explore the data through statistical analysis. Figure 4 shows a 3-dimensional scatterplot of the data, where X1, X2 and X3 correspond to the magazine circulation, percentage of male readership and median income, respectively.

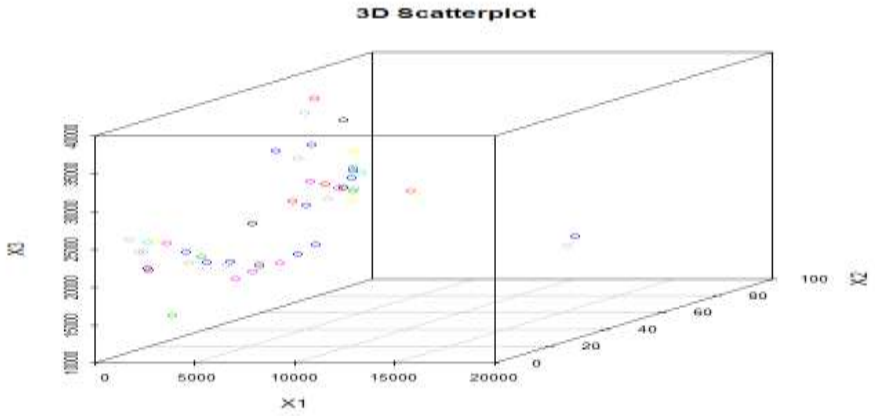

Fig. 4: 3D Scatterplot of Response Variables

The estimated ordinary least squares equation is

Mean Page Cost = -8642.5 + 5.3 Circulation + 11.0 Percent Male + 1.2 Median Income

I use the paired t-test to test the significance of each feature variable circulation, percentage of male readers, and median income of readers. For each feature variable, I start with the hypotheses that $H_0: \beta_1 = 0$ and $H_a: \beta_1 \neq 0$. The test statistic that will be used to test H_0 against H_a is given by $t_1^* = \frac{b_1 - 0}{\sqrt{\text{Var}(b_1)}}$. Using the statistical package R, I found that the observed value of this statistic for circulation is $t_{1,\text{obs}}^* = 9.958$. With a 95% confidence level, I have sufficient evidence to reject H_0 with a p-value of 0.00001 and smaller than $\alpha = 0.05$. The data provide sufficient evidence that there is a linear relationship between the response variable page cost and the predictor variable circulation, after adjusting for the predictor variable percentage of male readership and median income.

Following a similar approach, I conducted paired t-tests for percentage of male readers and median income of readers. I found the p-value to be 0.8874, greater than $\alpha = 0.05$, for percentage of male readers. Thus, the result of the test is not significant at the 5% significance level. Therefore, the data do not provide

sufficient evidence that there is a linear relationship between page cost and percentage of male readership, after adjusting for the predictor variable circulation and median income. For median income of readers, I found the p-value to be 0.0273, smaller than $\alpha = 0.05$. Thus, the result of the test is significant at the 5% significance level. The data provide sufficient evidence that there is a linear relationship between page cost and median income, after adjusting for the predictor variable circulation and percentage of male readership.

The paired t-tests show that circulation and median income of readers have a significant effect on advertising page cost in magazines. This is reasonable as greater circulation means more readership and popularity. Greater popularity means more demand, which gives the magazine more bargaining power to charge more for advertisements. Higher median income of readers for a magazine means that those readers have more purchasing power, and are likely to purchase more expensive items. As a result, those magazines can advertise more expensive products and charge the advertisers more.

The model's adjusted R^2 is 0.6734, meaning that 67.34% of the total variation in the dependent variable (page cost) is explained by all of the independent variables (circulation, percentage of male readership, and median income) taken together, after adjusting for the number of independent (predictor) variables. I used adjusted R^2 instead of R^2 (coefficient of determination) because R^2 tends to overestimate the strength of the association, especially for models with more than one independent variables. The equations for calculating R^2 and adjusted R^2 are below.

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

$$R^2_{adj} = 1 - \frac{\frac{SSE}{n-p}}{\frac{SST}{n-1}} = 1 - \frac{MSE}{\frac{SST}{n-1}}$$

In calculating R^2, SSE always decreases, so R^2 always increases. This is because the model will only add a variable if it reduces the SSE. In other words, a variable will only be added if it improves the model by reducing the error. On the other hand, in calculating the adjusted R^2, $\frac{SST}{n-1}$ does not depend on the number of independent variables in the model, but MSE does. The adjusted R^2 increases if the new variable reduces the MSE. It can also decrease if the new variable reduces the MSE, but not enough to offset the loss of one degree of freedom due to an extra regression coefficient. In summary the value of R^2 always increases with an additional independent variable, making it hard to interpret the strength of variability in the target variable that is explained by the new predictor variable. The adjusted R^2 is the more informative measure because it does not always increase with new variables added.

IV. CONCLUSION

This paper shows the use of a machine learning model, the ordinary least squares linear regression, to explore the effects of magazine circulation, gender and median income of readers on advertising page costs in magazines. The model is evaluated using paired t-tests and adjusted R^2. The adjusted R^2 is superior to the R^2 because it does not always increase with added variables. In other words, the adjusted R^2 can identify features that do not improve the model as well as expected. The results show that circulation and median income of readers have significant effect on advertising costs. A potential future extension would be to explore any new features that may affect advertising costs in magazines and explore any multicollinearity in the model that will increase the standard errors of the parameter estimates.

REFERENCES

[1]. James EL, Vanden Bergh BG. An Information Content Comparison of Magazine Ads across a Response Continuum from Direct Response to Institutional Advertising. Journal of Advertising. 1990;19(2):23-29.
[2]. Healey JS, Kassarjian HH. Advertising Substantiation and Advertiser Response: A Content Analysis of Magazine Advertisements. Journal of Marketing. 1983;47(1):107-117.
[3]. Fuchs DA. Two Source Effects in Magazine Advertising. Journal of Marketing Research. 1964;1(3):59-62.
[4]. Soley LC, Reid LN. Satisfaction with the Informational Value of Magazine and Television Advertising. Journal of Advertising. 1983;12(3);27-31
[5]. Morris LA, Brinberg D, Klimberg R, Rivera C, Millstein LG. The Attitudes of Consumers toward Direct Advertising of Prescription Drugs. Public Health Reports. 1986;101(1):82-89.

Andy W. Chen " A Machine Learning Analysis Of Advertising In Magazine." International Journal of Business and Management Invention (IJBMI) , vol. 07, no. 06, 2018, pp. 73–77.