Multiple Regression Case Study

Magazine Advertising

What factors influence the price of advertisements in magazines? Suppose you are part of a team of consultants hired by a retail clothing company wishing to place advertisements in at least one magazine. They are curious about what types of costs they can expect for magazines with different readership bases, so they most effectively utilize their advertising budget. Your team has collected a dataset of 44 consumer magazines and has found that the mean cost for a one-page advertisement is $82,386, but the standard deviation is $46,191.

What number should be used to best estimate the advertising costs?

Your team realizes that there may be many variables affecting cost of a one-page advertisement. You have augmented the original dataset of 44 magazines by measuring more characteristics of the magazines and their audiences that may be useful in understanding the one-page advertisement costs better. The variables are the following:

- Magazine Name
- Cost of a four-color, one-page ad
- Circulation (projected, in thousands)
- Percent male among the predicted readerships
- Median household income of readership

Your goal is to analyze the data using Multiple Linear Regression methods and choose the best model to explain the differences in advertising costs between the different titles and then to predict what the retail clothing company should expect to pay for advertising in the different magazines.

Answer the following questions:

1. Examine the variables and their relationships to each other:

a. First look at how each variable (all 4 of them) behaves on its own by creating histograms of each. Is there any apparent skewness in any of the graphs? Explain.

b. Now explore the linear relationship between page cost and each of the audience variables individually by constructing scatterplots of all three pairs. Do you see any strong relationships? Are they linear? Explain your answer.

2. Perform a Multiple Linear Regression analysis using all the audience variables AND perform a residual analysis using the graphs.

a. Is this multiple regression model useful? Provide statistical evidence to support your answer and where appropriate use a significance level of 5%.

b. What is the estimated regression equation?

c. Examine each of the audience variables individually to determine which are contributing significantly to the model. Which independent variables would you recommend keeping in the model? (Use a significance level of 5%.) [Note: do not eliminate any variable(s) at this stage.

d. Evaluate the regression assumptions of linearity and homoscedasticity (constant variance of the error term) by assessing your residual plot. Be specific about your evaluation and describe any suggestions you have for remedying any problems. [Hint: for suggestions you may read questions 3 and 4 below.]

e. Using this model with ALL the variables, provide a point estimate and an appropriate interval to the retail clothing company for the amount that they should expect to pay for a full-page ad in a magazine with a projected audience of 2,125,000 readers, 45 percent of which are male, with a median income of $50,000. Include notation and units. Interpret these results.

3. Often, when dealing with dependent variables that represent financial data (income, price, etc.), using the natural log of the dependent variable will help to alleviate problems that may be causing patterns in residuals/violations of the required conditions. Re-run the Multiple Regression analysis using the natural log of the page cost variable instead. Re-run the Regression using this new variable as the dependent variable against all 3 independent variables, again creating residual plot for this model.

a. Is this new multiple regression model useful? Provide statistical evidence to support your answer. Does the new Regression model seem better that the previous ones? Why or why not?

b. Examine each of the audience variables individually to determine which are contributing significantly to the new model. Which audience variables would you recommend keeping in the new model? [Note: do not eliminate any variable(s) at this stage.] How does this compare to the results in question 2?

c. Evaluate the regression assumptions of linearity and homoscedasticity by assessing your new residual plot. Be specific about your evaluation and describe any suggestions you have for remedying any problems. [Hint: for suggestions you may read question 4 below.]

4. Since you have switched to using the natural log of the Pagecost variable, you now need to recreate scatter plots using this as your dependent variable and each of the 3 independent variables on the x-axis (the result will be 3 separate scatterplots). The circulation variable has the most noticeable relationship to the natural log of Pagecost. This is a logarithmic type of relationship; to transform this curved relationship into a linear kind, a natural log transformation needs to be applied to the circulation variable. Do the transformation.

a. If natural log of independent variable – Circulation and median income is taken, and Page cost and percentage of men is not transformed, examine how the independent variables are now individually contributing to determine page cost.

b. Apply Z score analysis on all the independent variables and check if an outlier exists in your data and if found treat those outliers.

Executive Summary:

You are given the task of summarizing your findings for the board of directors of the retail clothing company. Since they are not all very well-versed in Regression techniques, you will need to explain things in easy-to-understand terms.

Within the summary, explain to them why you would recommend/not recommend using Linear Regression model to best forecast the cost of one-page advertisements.

Also, describe what this model indicates (very briefly) about the relationship between the page cost and your chosen variables.

PS – The Dataset has been created to highlight the various concepts of Regression and to check your understanding in it. It might be a possibility that your model is not performing well with the given Dataset. In this condition, please specify your understanding on the data and if you feel your final model does not sufficiently explain page cost, include your recommendations for improving it.