BUAD 310 Case 1 Solutions

(Total 100 points)

Due December 10, 2010

In this case you will apply statistical techniques learned in the Regression part of BUAD 310.

Instructions:

- **This exercise uses data from the file MagazineAds.MTW which you can download from Blackboard.**
- **The entire case should be typed in word document and clearly written and submitted with supporting documents (Minitab printouts). The printouts can be directly copied and pasted into the word document. You need to include the relevant graphs and numerical output so we can cross-check your answers with the output.**

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[1]:

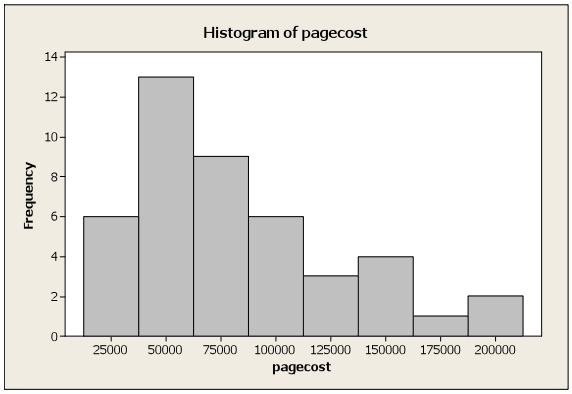
Magazine Name
Cost of a four-color, one-page ad
Circulation (projected, in thousands)
Percent male among the predicted readership
Median household income of readership

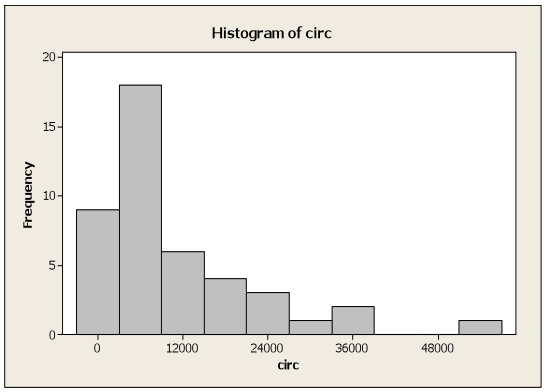
Your goal is to analyze the data with Minitab using Multiple Linear Regression methods and choose the best model to explain the differences in advertising costs between the different titles and then to predict what the retail clothing company should expect to pay for advertising in the different magazines.

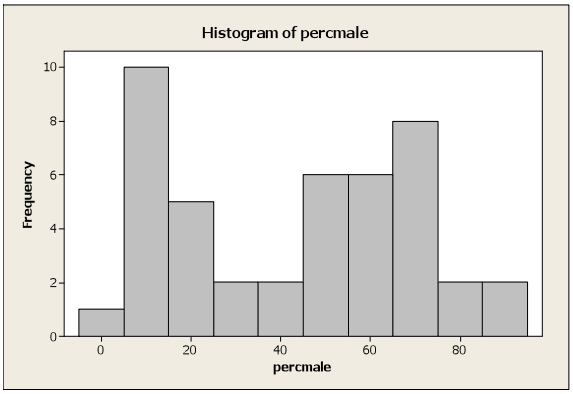
Answer the following questions:

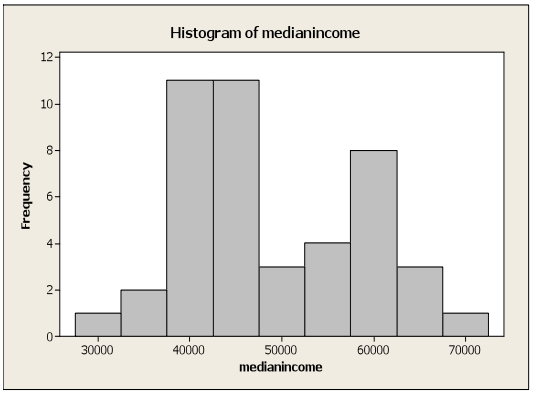
[1] Data are from *Mediamark Research Magazine Qualitative Audiences Report*, Spring 1996, and *SRDS Consumer Magazine Advertising Source*, July 1997, Volume 79 Number 7.

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. [Minitab: Graph → Histogram → Simple] Is there any apparent skewness in any of the graphs? Explain.



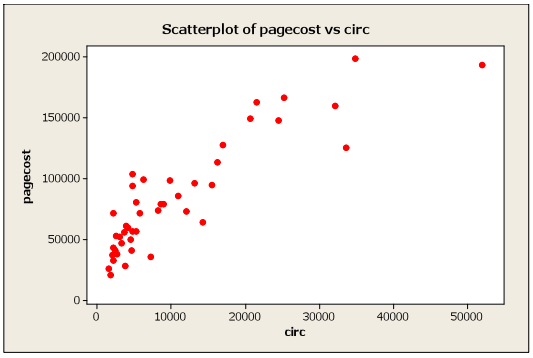


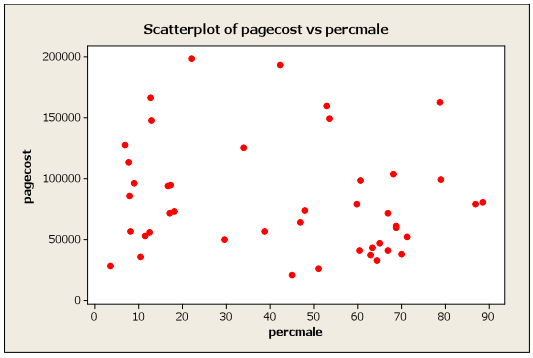


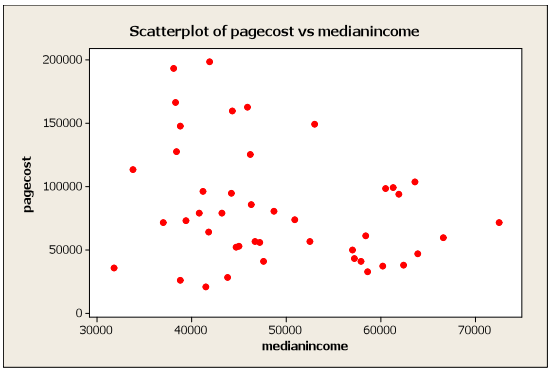


Pagecost and Circulation seem to be pretty right-skewed. Percmale and Medianincome seem to be bimodal and do not really have any pattern of skewness.

   b. Now explore the linear relationship between pagecost and each of the audience variables individually by constructing scatterplots of all three pairs. [Minitab: Graph → Scatterplot → Simple] Do you see any strong relationships? Are they linear? Explain your answer.







Pagecost and Circulation seem to have a moderately strong curved relationship (close to a natural log relationship). There seems to be no clear linear relationship between Pagecost and the other two audience variables.

2. Perform a Multiple Linear Regression analysis using all the audience variables AND perform a residual analysis using the graphs. [Minitab: Stat → Regression → Regression. After filling in the proper Response and Predictor variables, choose Graphs and check the box for *Residuals versus fits*.]

Regression Analysis: pagecost versus circ, percmale, medianincome

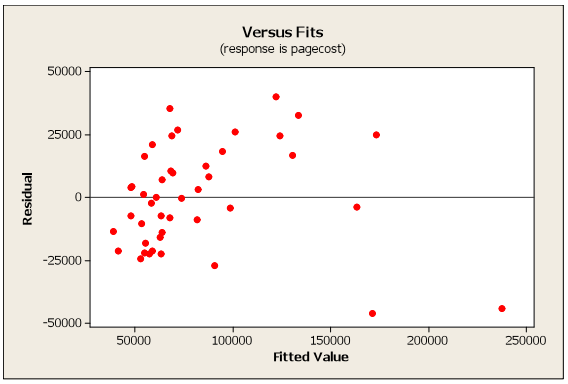
```
The regression equation is
pagecost = 7008 + 3.96 circ - 35 percmale + 0.698 medianincome

Predictor       Coef  SE Coef      T      P
Constant        7008    20383   0.34  0.733
circ          3.9557   0.3272  12.09  0.000
percmale       -34.7    147.6  -0.24  0.815
medianincome  0.6980   0.4253   1.64  0.109

S = 21350.7   R-Sq = 80.1%   R-Sq(adj) = 78.6%

Analysis of Variance

Source          DF           SS           MS      F      P
Regression       3  73510831804  24503610601  53.75  0.000
Residual Error  40  18234125601    455853140
Total           43  91744957405
```



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

P-value for the F-test is 0.000, so yes, the model is useful in explaining the response.

b. What is the estimated regression equation?

The regression equation is

pagecost = 7008 + 3.96 circ - 35 percmale + 0.698 medianincome

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.]

Circulation is very strongly significant (t-test p-value 0.000), Percmale is not significant (t-test p-value is 0.815), and Medianincome is also not significant (t-test p-value 0.109).

Since Circulation has p-value less than 0.05, I would keep it in the model. Percmale, however, should first be taken out since it is the most insignificant variable due to its highest p-value of 0.815. If necessary, I would remove Medianincome as well.

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.]

There is a certain pattern present in the residual plot, hence the linearity assumption is not satisfied: perhaps the relationship is not linear. Heteroscedasticity does not seem as evident here. A transformation, applied to the variables, may fix these problems. For example, I would probably try to transform the Pagecost and/or Circulation variables by taking the natural log.

e. Using this model with ALL the variables, provide a point estimate and an appropriate 95% 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.

New Obs	Fit	SE Fit	95% CI	95% PI
1	48753	4199	(40267, 57238)	(4775, 92731)

Point estimate: $48,753. Since this is for a *particular* magazine, I would use the 95% prediction interval: ($4,775, $92,731).

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. [Minitab: Calc → Calculator, then enter "Lncost" in the *Store Result in Variable*. In *Expression*, enter "LN(pagecost)" and click *OK*. This will create a new variable that will be the *natural log* of the page cost variable.]. Re-run the Regression using this new variable as the dependent variable against all 3 independent variables, again creating residual plot for this model.

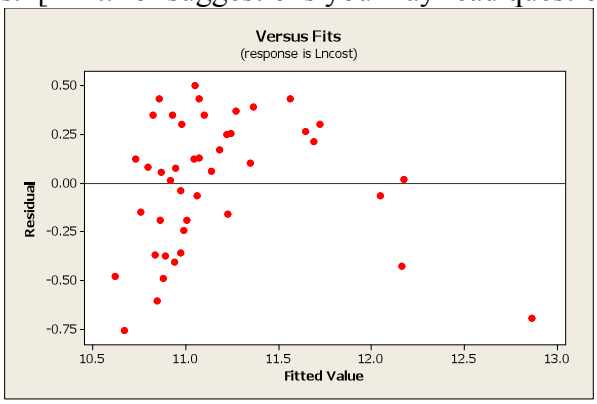
   a. Is this new multiple regression model useful? Provide statistical evidence to support your answer and where appropriate use a significance level of 5%. Does the new Regression model seem better that the previous ones? Why or why not?

P-value for the F-test is 0.000, so yes, the model is useful in explaining the response. The model has a R-square (R^2) of 65.2%, which is lower than the one for the first model 80.1%, so the second model does not fit the data as well.

   b. Examine each of the audience variables individually to determine which are contributing significantly to the new model. Use a significance level of 5%. 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?

The results are very similar to those of the first model: Circulation is very strongly significant (t-test p-value 0.000), Percmale is not significant (t-test p-value is 0.582), and Medianincome is not significant either (p-value 0.077). So I would keep Circulation in the model.

   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.]



Homoscedastisity assumption does not seem to be clearly violated. There is a certain pattern present in the residual plot, hence the linearity assumption is not satisfied: perhaps the relationship is not linear. A transformation, applied to the variables, may fix this problem. I would suggest transforming one of the X variables due to the potential non-linearity.

4. Since you have switched to using the *natural log* of the Pagecost variable, you now need to re-create scatterplots 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 in Minitab. [The steps are described in part 3. This will create a new variable that will be the transformed version of the original circulation variable.] Re-run the Regression and residual analysis using the new variable in place of the old one (leave the dependent variable as *natural log* of Pagecost and leave the other two audience variables as they are).

Regression Analysis: lnCost versus lnCirc, percmale, medianincome

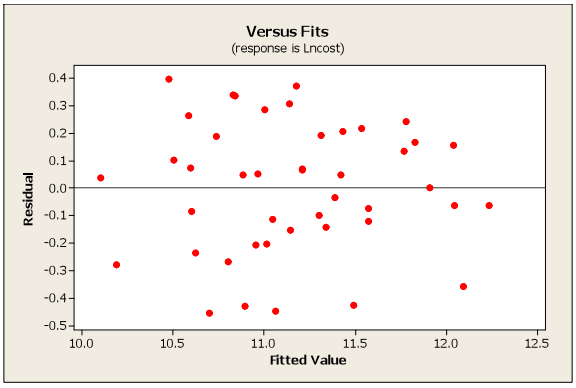
```
The regression equation is
Lncost = 4.85 + 0.622 lnCirc + 0.00035 percmale + 0.000016 medianincome

Predictor           Coef     SE Coef      T      P
Constant          4.8488      0.5411   8.96  0.000
lnCirc           0.62161     0.04597  13.52  0.000
percmale        0.000347    0.001690   0.21  0.838
medianincome  0.00001632  0.00000488   3.34  0.002

S = 0.244025   R-Sq = 82.7%   R-Sq(adj) = 81.5%

Analysis of Variance

Source           DF       SS      MS      F      P
Regression        3  11.4263  3.8088  63.96  0.000
Residual Error   40   2.3819  0.0595
Total            43  13.8083
```



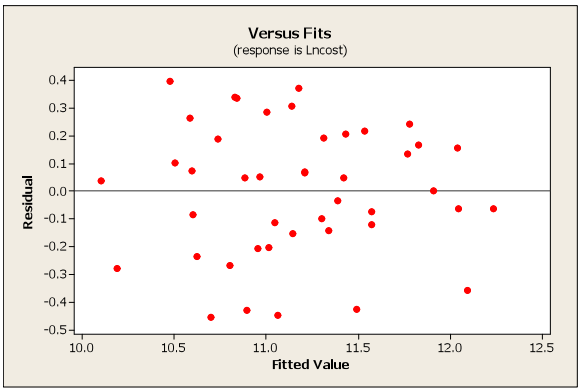
a. Has the usefulness of the model changed? Is this model better or worse than the previous model? Support your answer.

P-value for the F-test is 0.000, so yes, the model is useful in explaining the response. The model has a R-square of 82.7%, the highest of the three models (80.1% and 65.2%), which means this model fits better than the other two.

b. Examine each of the independent variables individually to determine which are contributing significantly to the newest model. Use a significance level of 5%. Which variables would you recommend keeping in the model? How does this compare to the prior results?

The results are different from the other two models. The significance of the variables has somewhat increased: Ln(Circulation) and Median income are very strongly significant (t-test p-values 0.000 and 0.002), Percmale is not significant (t-test p-value is 0.838).

c. Evaluate the regression assumptions of linearity and homoscedasticity by assessing your new residual plot. Briefly comment on the new results.



The residual plot looks fine now, none of the assumptions are violated. The log transformations have fixed the problems!

d. Finally, remove any variables that you deem to be insignificant and re-run the model. Using this model, provide a point estimate and an appropriate 95% interval to the retail clothing company client for the amount that they should expect to pay <u>on average</u> if they were to place <u>many</u> ads in magazines with the following characteristics: projected circulation of 2,125,000 readers, 45 percent male, and median income of $50,000. Include notation and units. Interpret your results.

Percmale should be removed as the only insignificant variable. Using the new model, the point estimate for the amount is exp(10.4431)=$34,306.84. (Make <u>sure you remembered to plug in the natural log of circulation, 7.6615, instead of 2125, and that you exponentiated the value</u>.) I would use the Confidence Interval, but the endpoints need to be exponentiated to get back to the actual predicted pagecosts. The lower endpoint of the 95% CI is exp(10.3146)=$30,169.90 and the upper endpoint is exp(10.5715)=$39,007.14. Thus, with 95% confidence, the amount they should expect to pay on average is between $30,169.90 and $39,007.14. (Relevant Minitab output is below).

Regression Analysis: lnCost versus lnCirc, medianincome

```
The regression equation is
```

```
Lncost = 4.84 + 0.621 lnCirc + 0.000017 medianincome

Predictor           Coef     SE Coef      T      P
Constant          4.8445      0.5343   9.07  0.000
lnCirc           0.62109     0.04536  13.69  0.000
medianincome  0.00001680  0.00000422   3.98  0.000

S = 0.241158   R-Sq = 82.7%   R-Sq(adj) = 81.9%

Analysis of Variance

Source          DF       SS      MS      F      P
Regression       2  11.4238  5.7119  98.21  0.000
Residual Error  41   2.3844  0.0582
Total           43  13.8083

Predicted Values for New Observations

New
Obs      Fit  SE Fit        95% CI              95% PI
  1  10.4431  0.0636  (10.3146, 10.5715)  (9.9394, 10.9467)

Values of Predictors for New Observations

New
Obs  lnCirc  medianincome
  1    7.66         50000
```

Executive Summary: (1 page)

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 which model and estimates you would recommend to best forecast the cost of one-page advertisements. Also, describe what this model indicates (very briefly) about the relationship between the pagecost and your chosen variables. If you feel your final model does not sufficiently explain pagecost, include your recommendations for improving it.