Homework #9 – Solutions

COR1-GB.1305 – Statistics and Data Analysis

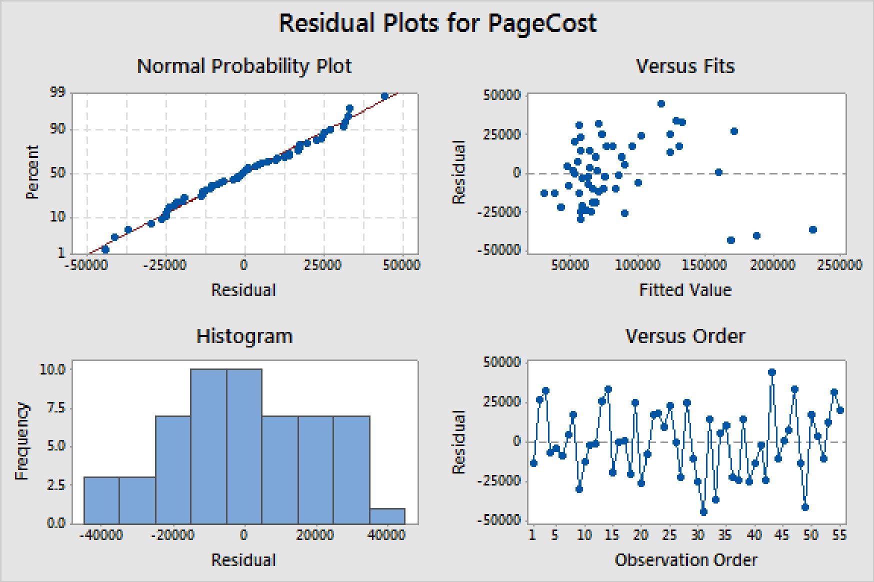
Problem 1

The file `Magazine.CSV` *contains data on advertising costs and characteristics of magazines. The response variable is* `PageCost`*, which represents the cost of a full-page color ad in the magazine.* `Circ` *is the circulation of the magazine (in thousands),* `MedIncome` *is the median income of the readers, and* `%Male` *is the percentage of the readers who are male. The square root of the circulation is given in* `SqrtCirc`*.*

In this problem, we will fit a regression model to predict the mean page cost of a magazine with `Circ` *of 10000,* `MedIncome` *of $40,000, and* `%Male` *of 50. We will not necessarily use all three predictors, only the ones that are useful for predicting* `PageCost`*.*

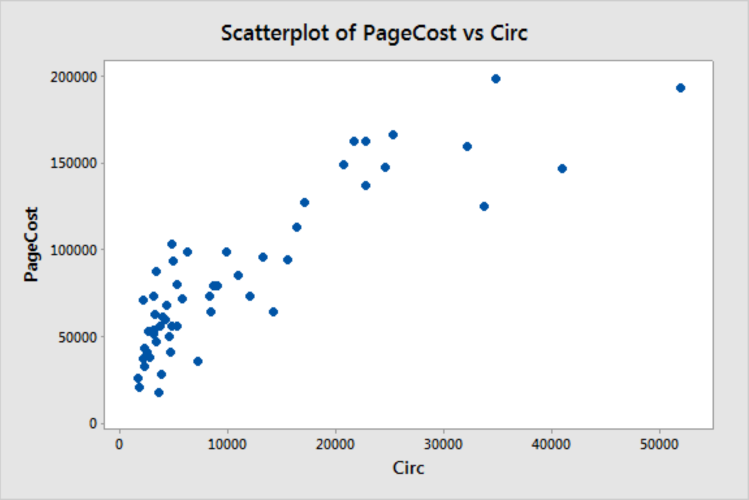
(a) *First, we will find an appropriate set of predictor variables.*

(i) *Run a multiple regression of* `PageCost` *on the original predictor variables (*`Circ`*,* `MedIncome` *and* `%Male`*). Before running it, click on Graphs, and check the box for Residuals plots: Four in one. Note that the residuals versus fit plot shows structure: a generally upward-sloping pattern, with three outliers at the right dragging things down. Identify the Magazines corresponding to the three outliers (all of which have a very large circulation).*



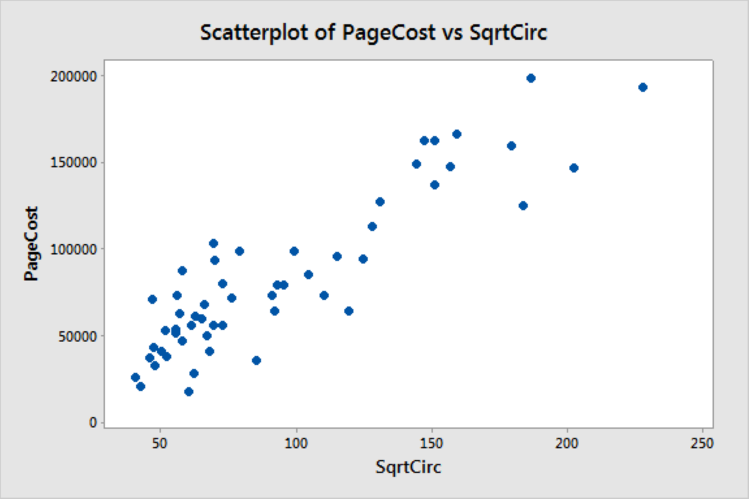
The three outliers are *People*, *Reader's Digest*, and *TV Guide*

(ii) *To investigate further, generate a scatterplot of* `PageCost` *versus* `Circ`*. Note that the plot is "bunched up" at the left, and "stretched out" at the right, and also a bit curved. In what way do the points identified as outliers in (a) deviate from the pattern in the plot here?*



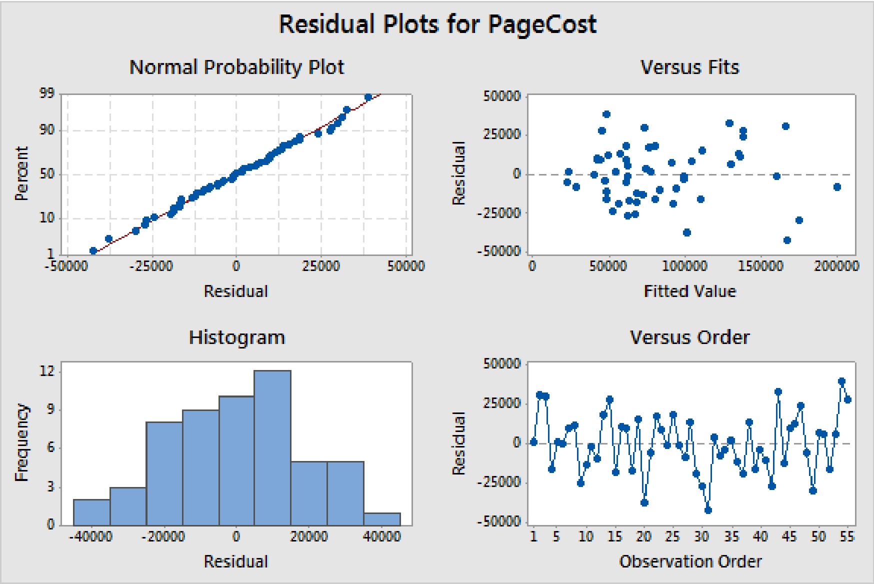
The three points are below the trend consistent with the other points. Also, the value of `Circ` is higher than for most of the other points.

(iii) *To try to improve the linear relationship, let's try working with the square root of Circulation (`SqrtCirc`) rather than the circulation itself. Plot `PageCost` versus `SqrtCirc`. Based on the plot, explain why it seems more appropriate to use `SqrtCirc` as an explanatory variable in a linear regression rather than `Circ`.*



The points seem more evenly spaced and the trend is no longer curved when we use `SqrtCirc`.

(iv) *Now, run a multiple regression of `PageCost` on `SqrtCirc`, `MedIncome` and `%Male`. Plot the residuals versus fitted values. Does it look better than in (i)?*



```
Analysis of Variance

Source        DF       Adj SS       Adj MS  F-Value  P-Value
Regression     3  93408864150  31136288050    87.65    0.000
  SqrtCirc     1  90285261164  90285261164   254.15    0.000
  MedIncome    1   3178783031   3178783031     8.95    0.004
  %Male        1    117077172    117077172     0.33    0.568
Error         51  18117098203    355237220
Total         54  1.11526E+11

Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
18847.7  83.76%     82.80%      81.04%

Coefficients

Term         Coef  SE Coef  T-Value  P-Value   VIF
Constant   -48399    16510    -2.93    0.005
SqrtCirc    945.6     59.3    15.94    0.000  1.16
MedIncome   0.966    0.323     2.99    0.004  1.66
%Male         -69      120    -0.57    0.568  1.47

Regression Equation

PageCost = -48399 + 945.6 SqrtCirc + 0.966 MedIncome - 69 %Male
```

The residuals versus fitted values plot looks better–the mean residual looks zero for all fitted values (before, the mean was negative for small fitted values, positive for medium fitted values, and negative for large fitted values).

(b) *Next we will investigate the regression model of `PageCost` on `SqrtCirc`, `MedIncome` and `%Male`.*

(i) *Based on the p-value for the Analysis of Variance F test for this model, does the regression seem to be useful for predicting `PageCost`? Does this mean that all variables are useful?*

The F-statistic p-value is reported as 0.000, which means that it is less than 0.001. Since the p-values is below 0.05, the regression seems useful for predicting Y. This does not mean that all variables are useful, only that at least one variable is useful.

(ii) *Which coefficients in the regression are statistically significant?*

The coefficients for `SqrtCirc` and `MedIncome` are statistically significant.

(iii) *Based on the p-values for the regression coefficients, which variables seem to be useless for predicting `PageCost`?*

The `%Male` variable seems useless for predicting `PageCost` after adjusting for the other variables.

(c) *Now, we will try to simplify the model by deleting useless predictors. Re-run the regression for* ***PageCost****, this time with just the two significant explanatory variables you found in part (b).*

```
Analysis of Variance

Source       DF       Adj SS       Adj MS  F-Value  P-Value
Regression    2  93291786978  46645893489   133.02    0.000
  SqrtCirc    1  90193993387  90193993387   257.21    0.000
  MedIncome   1   3635251757   3635251757    10.37    0.002
Error        52  18234175375    350657219
Total        54  1.11526E+11

Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
18725.8  83.65%     83.02%      81.54%

Coefficients

Term         Coef  SE Coef  T-Value  P-Value   VIF
Constant   -46126    15925    -2.90    0.006
SqrtCirc    943.8     58.8    16.04    0.000  1.16
MedIncome   0.865    0.269     3.22    0.002  1.16

Regression Equation

PageCost = -46126 + 943.8 SqrtCirc + 0.865 MedIncome
```

(i) *Did the R^2 go down by much compared to the regression in (b)? Is the F-statistic still significant? What does this suggest about the deleted predictor variable?*

The R^2 went from to 83.76% to 83.65%. The F-statistic is still significant. This suggests that deleting `%Male` did not affect the predictive performance of the model.

(ii) *Are the coefficients of both variables statistically significant?*

The coefficients of both variables are statistically significant.

(d) *Finally, we will use the simplified multiple regression model to predict* `PageCost`*.*

(i) *Get a 95% confidence interval for the mean page cost of a magazine with a* `SqrtCirc` *of 100, and a median income of $40,000. To do this, after running the regression click on* Stat ⇒ Regression ⇒ Regression ⇒ Predict*. Then, enter 100 in the first line under* `SqrtCirc` *and enter 40000 in the first line under* `MedIncome`*.*

```
Regression Equation

PageCost = -46126 + 943.8 SqrtCirc + 0.865MedIncome

Variable   Setting
SqrtCirc       100
MedIncome    40000

    Fit   SE Fit         95% CI              95% PI
82839.7  3197.96  (76422.5, 89256.8)  (44719.5, 120960)
```

The 95% confidence interval is: $(76422.5, 89256.8)$. With 95% confidence we can say that the mean page cost of a magazine with circulation $(100)^2$ and median income $40,000 is between $76,422.50 and $89,256.80.

(ii) *Report the 95% prediction interval.*

The 95% prediction interval is: $(44719.5, 120960)$. In the fitted model, 95% of page costs for magazines with circulations of $(100)^2$ and median incomes of $40,000 are between $44,719.50 and $120,960.00.

(iii) *What is the difference between the prediction interval and the confidence interval?*

The prediction interval is for the *all prices* of magazines with the given values of the predictor variables; the confidence interval is for the *mean price* of magazines with the given values of the predictor variables.