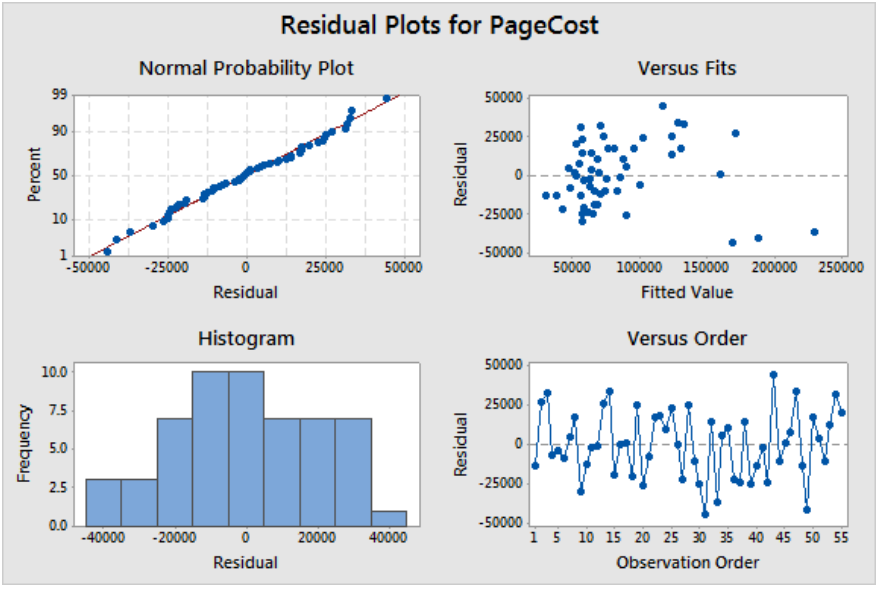
Homework #9 – Solutions
COR1-GB.1305 – Statistics and Data Analysis

Problem 1

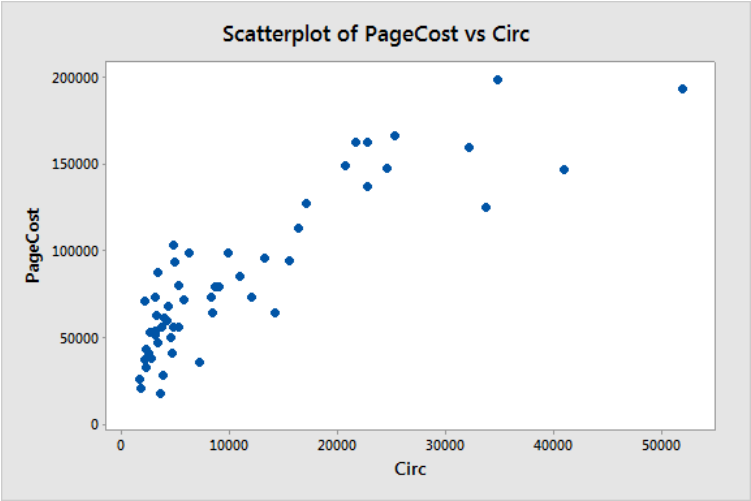
The file `Magazine.CSV` *contains data on advertising costs and characteristics of magazines. The response variable is* `PageCost`*, which represents the cost of a full-page color ad in the magazine.* `Circ` *is the circulation of the magazine (in thousands),* `MedIncome` *is the median income of the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

(a) *Run a multiple regression of* `PageCost` *on* `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).*



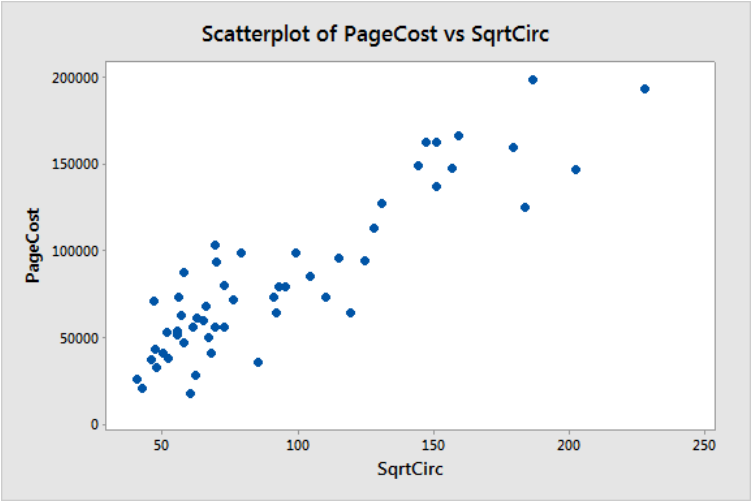
The three outliers are *People*, *Reader's Digest*, and *TV Guide*

(b) *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?*



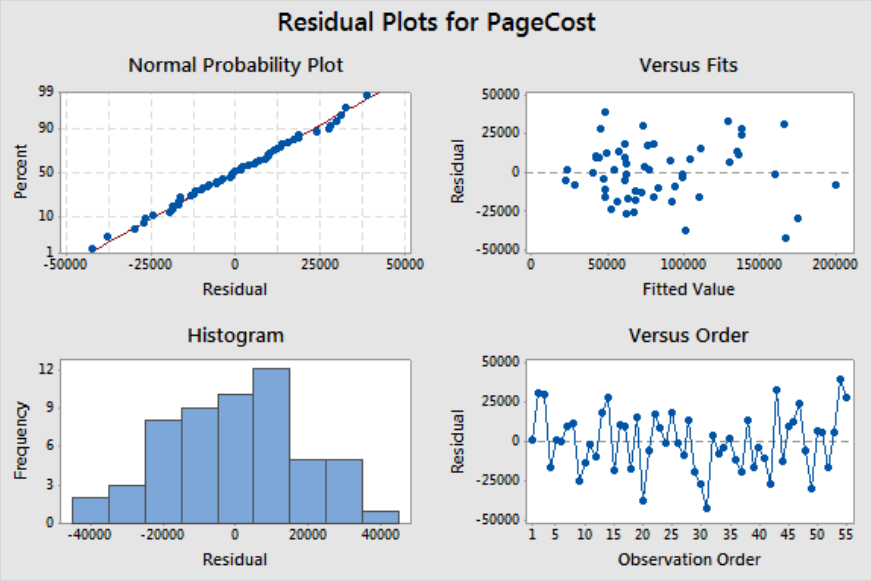
The three points are below the trend consistent with the other points. Also, the value of `Circ` is higher than for most of the other points.

(c) *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`.

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




```
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
```

Which coefficients in the regression are statistically significant? Based on the p-values for the regression coefficients, which variables seem to be useless for predicting `PageCost`*?*

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).

The coefficients for `SqrtCirc` and `MedIncome` are statistical significant. The `%Male` variable seems useless for predicting `PageCost` after adjusting for the other variables.

(e) *The F-statistic and corresponding p-value in the Analysis of Variance part of the output provides a test of the null hypothesis that the regression is useless for predicting Y, i.e., that all regression parameters besides the intercept are zero. Based on the p-value, does the regression seem to be useful for predicting Y? Does this mean that all variables are useful? (Remember your answer to part (d).*

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.

(f) *Re-run the regression for* `PageCost`*, this time with just the two explanatory variables* `SqrtCirc` *and* `MedIncome`*. Are the coefficients of both variables statistically significant?*

```
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
```

The coefficients of both variables are statistically significant.

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 ⇒ Preict*. 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)
```

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.

Did the R^2 go down by much compared to the regression in (d)? Is the F-statistic still significant? What does this suggest about the deletion of the `%Male` *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.

(g) *Run a simple regression and scatterplot of* `PageCost` *versus* `MedIncome`*. Is the coefficient of* `MedIncome` *now statistically significant? Why is this puzzling in view of the regression output in (f)? Remember, however, that the meaning and interpretation of the coefficient of a given variable depend on what other variables are included in the model.*

```
Analysis of Variance

Source       DF       Adj SS      Adj MS  F-Value  P-Value
Regression    1   3097793590  3097793590     1.51    0.224
  MedIncome   1   3097793590  3097793590     1.51    0.224
Error        53  1.08428E+11  2045814505
Total        54  1.11526E+11

Model Summary

      S   R-sq  R-sq(adj)  R-sq(pred)
45230.7  2.78%      0.94%       0.00%

Coefficients

Term         Coef  SE Coef  T-Value  P-Value   VIF
Constant   118874    29360     4.05    0.000
MedIncome  -0.741    0.602    -1.23    0.224  1.00

Regression Equation

PageCost = 118874 - 0.741 MedIncome
```

The coefficient of `MedIncome` is no longer statistically significant. This might be puzzling, because the coefficient was significant when `SqrtCirc` was in the model. Apparently, `MedIncome` is only useful as a predictor after adjusting for `SqrtCirc`; it is not useful by itself.

. . . . . .

Problem 2

Consider `DiamondPrices.CSV`*, which you already studied in the context of simple regression.* `Clarity`*,* `Code` *and* `Color` *are categorical variables, which are numerically coded in* `ClarityCode`*,* `ColorCode` *and* `CutCode`*. (For example, Good, Very Good and Ideal cuts receive* `Cut Codes` *of 1, 2, 3, respectively.) Even though these are ordinal/categorical variables, please enter them in Minitab as "Continuous Predictors", since we will treat them as numerical variables.*

(a) *Run a multiple regression of* `Price` *on* `ClarityCode` *and* `CutCode`*. Based on the output, do these explanatory variables seem useful for predicting* `Price`*?*

```
Analysis of Variance

Source          DF      Adj SS    Adj MS  F-Value  P-Value
Regression       2      237352    118676     0.02    0.981
  ClarityCode    1      163141    163141     0.03    0.873
  CutCode        1       83414     83414     0.01    0.909
Error          614  3900089596   6351937
  Lack-of-Fit   18   255492373  14194021     2.32    0.002
  Pure Error   596  3644597223   6115096
Total          616  3900326948

Model Summary

      S   R-sq  R-sq(adj)  R-sq(pred)
2520.31  0.01%      0.00%       0.00%

Coefficients

Term          Coef  SE Coef  T-Value  P-Value   VIF
Constant      4564      485     9.40    0.000
ClarityCode  -11.0     68.6    -0.16    0.873  1.00
CutCode         19      167     0.11    0.909  1.00

Regression Equation

Price = 4564 - 11.0 ClarityCode +19 CutCode
```

The p-value for the F test is 0.981. Based on this, the explanatory variables do not seem useful for predicting `Price`.

(b) *Run a multiple regression of* `Price` *on* `ClarityCode`*,* `ColorCode` *and* `CutCode`*. In what way do p-values for the individual coefficients and the F-statistic seem to provide contradictory evidence on whether any of these variables is helpful for predicting* `Price` *in the given model? (As we discussed in class, the F-statistic is the best place to go to make a decision on this question.) Does the R^2 in this regression suggest that these three variables by themselves are very useful for predicting* `Price`*?*

```
Analysis of Variance

Source          DF      Adj SS    Adj MS  F-Value  P-Value
Regression       3    37633773  12544591     1.99    0.114
  ClarityCode    1     2281584   2281584     0.36    0.548
  ColorCode      1    37396421  37396421     5.93    0.015
  CutCode        1      545565    545565     0.09    0.769
Error          613  3862693175   6301294
  Lack-of-Fit  106   994565287   9382691     1.66    0.000
  Pure Error   507  2868127887   5657057
Total          616  3900326948

Model Summary

      S   R-sq  R-sq(adj)  R-sq(pred)
2510.24  0.96%      0.48%       0.00%

Coefficients

Term          Coef  SE Coef  T-Value  P-Value   VIF
Constant      3928      549     7.15    0.000
ClarityCode  -41.8     69.5    -0.60    0.548  1.04
ColorCode    161.1     66.1     2.44    0.015  1.04
CutCode         49      167     0.29    0.769  1.01

Regression Equation

Price = 3928 - 41.8 ClarityCode + 161.1 ColorCode + 49 CutCode
```

The apparent contradiction here is that the F test does not indicate that the model is useful ($p = 0.114$), but the t test on `ColorCode` indicates a significant coefficient ($p = 0.015$). The R^2 is 0.96%, which is extremely low. Both the F test and the R^2 indicate that these three variables by themselves are not very useful for predicting `Price`.

(c) *Next, run a multiple regression of* `Price` *on* `Carats`, `ClarityCode`, `ColorCode` *and* `CutCode`. *What happened to the variables that had insignificant coefficients in the regression in (b)? And what happened to the* R^2*?*

```
Analysis of Variance

Source          DF      Adj SS      Adj MS  F-Value  P-Value
Regression       4  3510752974   877688243  1378.80    0.000
  Carats         1  3473119201  3473119201  5456.09    0.000
  ClarityCode    1   213882083   213882083   336.00    0.000
  ColorCode      1   263076238   263076238   413.28    0.000
  CutCode        1     5830788     5830788     9.16    0.003
Error          612   389573974      636559
  Lack-of-Fit  549   375941972      684776     3.16    0.000
  Pure Error    63    13632003      216381
Total          616  3900326948

Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
797.846  90.01%     89.95%      89.83%

Coefficients

Term          Coef  SE Coef  T-Value  P-Value   VIF
Constant     -7132      230   -31.00    0.000
Carats        9377      127    73.87    0.000  1.13
ClarityCode  420.7     23.0    18.33    0.000  1.12
ColorCode    433.9     21.3    20.33    0.000  1.07
CutCode      160.6     53.1     3.03    0.003  1.01

Regression Equation

Price = -7132 + 9377 Carats + 420.7 ClarityCode + 433.9 ColorCode + 160.6 CutCode
```

All of the variables have significant coefficients. Also, the R^2 is dramatically higher (it is now 90.01%).

(d) *For the regression in part (c), do the p-values for the individual coefficients and the F-statistic have any apparent contradiction?*

There is no apparent contradiction. The tests on the individual coefficients indicate that all are significant. The F statistic indicates that the model is useful (at least one coefficient is nonzero).

(e) *Using the regression model in part (c), predict the retail price of a diamond weighing* 0.6 *carats with a VS2 clarity (*`ClarityCode` $= 3$*), a G color (*`ColorCode` $= 4$*) and a Very Good cut (*`CutCode` $= 2$*).*

```
Regression Equation

Price = -7132 + 9377 Carats + 420.7 ClarityCode + 433.9 ColorCode + 160.6 CutCode

Variable     Setting
Carats           0.6
ClarityCode        3
ColorCode          4
CutCode            2

    Fit   SE Fit        95% CI              95% PI
1813.20  58.0415  (1699.21, 1927.18)  (242.207, 3384.19)
```

The model predicts the price to be \$1,813.20. More precisely, according to the prediction interval, typical diamonds with these characteristics have prices between \$242.20 and \$3,384.19. According to the confidence interval, we are 95% confident that average price of all diamonds with these characteristics is between \$1,699.21 and \$1,927.18.

(f) *Run a simple regression of* `Price` *on* `Carats`*, and note the* R^2*. Comparing this with the results from (c), do you think that the* `ClarityCode`*,* `ColorCode` *and* `CutCode` *(taken together) are worthwhile for predicting* `Price`*, above and beyond what can be obtained using* `Carats` *alone?*

```
Analysis of Variance

Source          DF      Adj SS      Adj MS  F-Value  P-Value
Regression       1  2962746890  2962746890  1943.40    0.000
  Carats         1  2962746890  2962746890  1943.40    0.000
Error          615   937580058     1524520
  Lack-of-Fit   99   263117475     2657752     2.03    0.000
  Pure Error   516   674462582     1307098
Total          616  3900326948


Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
1234.71  75.96%     75.92%      75.80%


Coefficients

Term       Coef  SE Coef  T-Value  P-Value   VIF
Constant  -2541      169   -15.04    0.000
Carats     8130      184    44.08    0.000  1.00


Regression Equation

Price = -2541 + 8130 Carats
```

For the simple regression model, we have $R^2 = 75.96\%$. For the multiple regression model, we had $R^2 = 90.01\%$ This is a big increase. It seems like the `ClarityCode`, `ColorCode` and `CutCode` (taken together) are worthwhile for predicting `Price` above and beyond what can be obtained using `Carats` alone.

.....