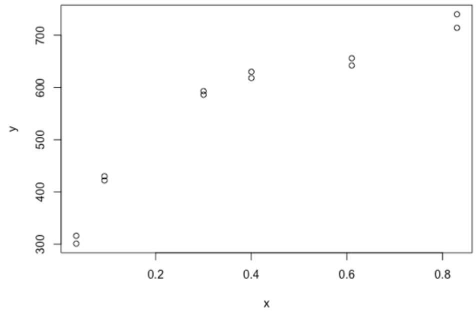
Name: CHAN King Yeung
SID: 1155119394
STAT3008 Assignment 4

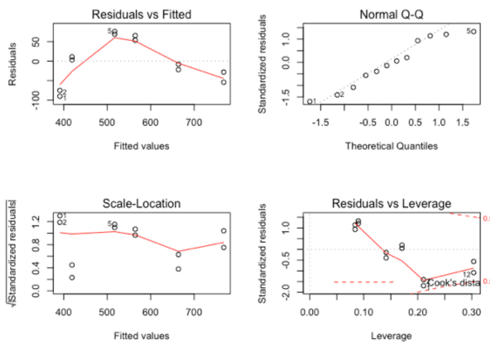
Question 2

a) The scatterplot shows a positive relationship between the RV and the EV. The data points do not do not follow a straight line, and it is less believe to be a linear relationship among 2 variables.



b) After executed the R code, it returns
$\hat{y}_i = 375.0036 + 473.7442x_i$
$R^2 = 0.8533076$

c) From the residuals and fitted values plot (first-row first graph), there is a trend on the residuals. The mean level is not zero and the variance is variable, which violate the "constant mean" and "constant variance" assumptions of the null plot. From the residuals and leverage graph (second-row second graph), there a few points far from the left. They are highly believed to be leverage points, and those points are separated from the data centre which violates "no separated points" of the null plot assumption.



d) After executed the R code, it returns

```
Influence measures of
         lm(formula = y ~ x) :

     dfb.1_   dfb.x   dffit cov.r   cook.d    hat inf
1  -0.98261  0.7647 -0.9836 0.795 0.383169 0.2107
2  -0.77291  0.6015 -0.7737 1.002 0.266220 0.2107
3   0.08572 -0.0619  0.0866 1.477 0.004145 0.1707
4   0.02298 -0.0166  0.0232 1.488 0.000299 0.1707
5   0.34392 -0.1176  0.4365 0.920 0.087157 0.0899
6   0.30647 -0.1048  0.3889 0.990 0.071817 0.0899
7   0.18550  0.0280  0.3520 1.018 0.059828 0.0839
8   0.14786  0.0223  0.2806 1.123 0.039914 0.0839
9   0.00338 -0.0356 -0.0555 1.432 0.001708 0.1414
10  0.00936 -0.0986 -0.1538 1.393 0.012941 0.1414
11  0.13490 -0.3062 -0.3595 1.661 0.069508 0.3035   *
12  0.27155 -0.6164 -0.7237 1.380 0.256665 0.3035
```

For every observation i, $|DFFITS_i| < 1$, $|DFBETA_i| < 1$ and $D_i < 1$ which suggest there is no influential point based on the DFFITS, the DFBETAS and the Cook's Distance; and
For every observation i, $h_{ii} < \frac{1}{3}$ which suggests there is no leveraged point on the data set

Question 3

a) After executed the R code, it returns

$\hat{\alpha}_0 = 1.58036341$
$\hat{\alpha}_1 = 0.41606845$
$\hat{\alpha}_2 = 0.06555584$
$RSS = 5814.13$

b) After executed the R code, it returns

```
Start:  AIC=435.96
y ~ 1

       Df Sum of Sq   RSS    AIC
+ xsq   1     62080  5869 286.12
+ x     1     59639  8310 307.68
<none>               67949 435.96
```

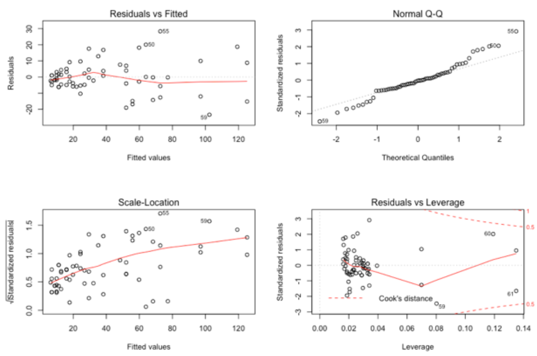
```
Step:  AIC=286.12
y ~ xsq

       Df Sum of Sq    RSS    AIC
<none>               5869.2 286.12
+ x     1    55.102 5814.1 287.54
```

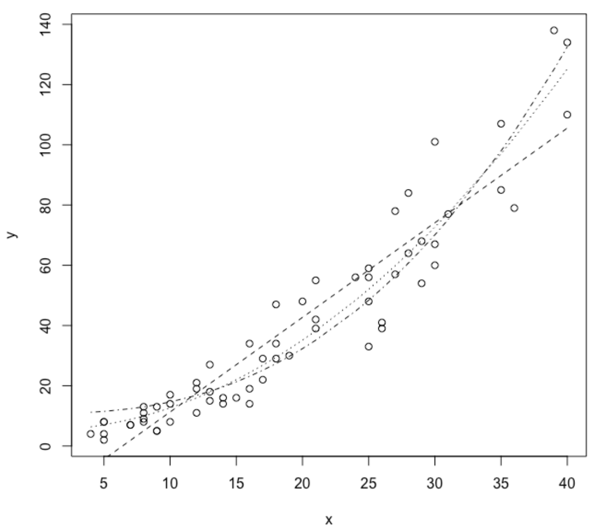
$\therefore$ the parsimonious model under AIC forward selection is $y_i = \alpha_0^* + \alpha_2^* x_i^2$

$\hat{\alpha}_0^* = 5.1347681$
$\hat{\alpha}_1^* = 0.0750361$
$RSS = 5869.232$

c) From the residuals and fitted values graphs (first-row first graph), the majority of data concentrate in the left hand side and the remaining data points far away from each other which violate “constant variance” assumption of the null plot. Since there is not a trend in the plot, “constant mean” assumption seems to be valid for the null plot. From the residuals and leverage graph (second-row second graph), however, there are a few points above and below ± 2 in the y-axis. Those data points are treated as outlier and those outliers violate the “no separated points” assumption of the null plot.



d) After executed the R code, it returns



e) After executed the R code, it returns

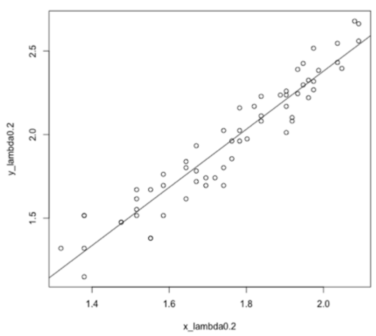
$RSS_{\lambda=1} = 8310.166$
$RSS_{\lambda=2} = 5869.232$
$RSS_{\lambda=2.5} = 6756.696$

i) Since $RSS_{\lambda=2}$ is the smallest RSS among the 3 RSS's, the transformation with $\lambda = 2$ is the best among the 3.

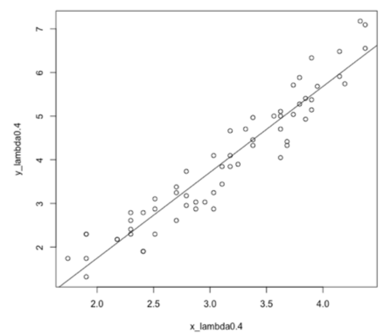
ii) The regression model of transformation with $\lambda = 2$ is a linear transformation of the model P. That is $\frac{x_i^2-1}{2}$ is the linear transformed term of x_i^2 in the model P. Since linear transformation of x_i^2 does not alter the column space of $\boldsymbol{X}$, the corresponding RSS remains unchanged after transformation. Thus, $RSS_{model\ P} = RSS_{\lambda=2}$.

f) After executed the R code, it returns

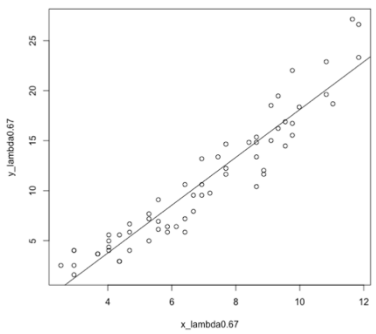
For $\lambda = 0.2$,



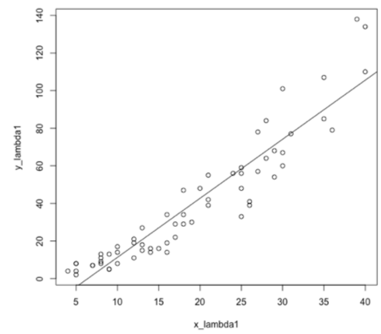
For $\lambda = 0.4$,



For $\lambda = 0.67$,



For $\lambda = 1$,



From the above scatterplots, the one with $\lambda = 0.2$ shows that more data points are close to the regression after transformation, which the regression can explain more variation of data. The number of leverage should be greatly reduced while comparing to other transformations. Moreover, with the following statistics

$RSS_{\lambda=0.2} = 0.7822444$
$RSS_{\lambda=0.4} = 12.06266$
$RSS_{\lambda=0.67} = 256.3591$
$RSS_{\lambda=1} = 8310.166$

The data set after transformed with $\lambda = 0.2$ has the smallest RSS among 4 transformations, which means the impact of leverage points is minimised with $\lambda = 0.2$ while comparing to other transformations. Thus, $\lambda = 0.2$ is able to provide the smallest number of leverage points.