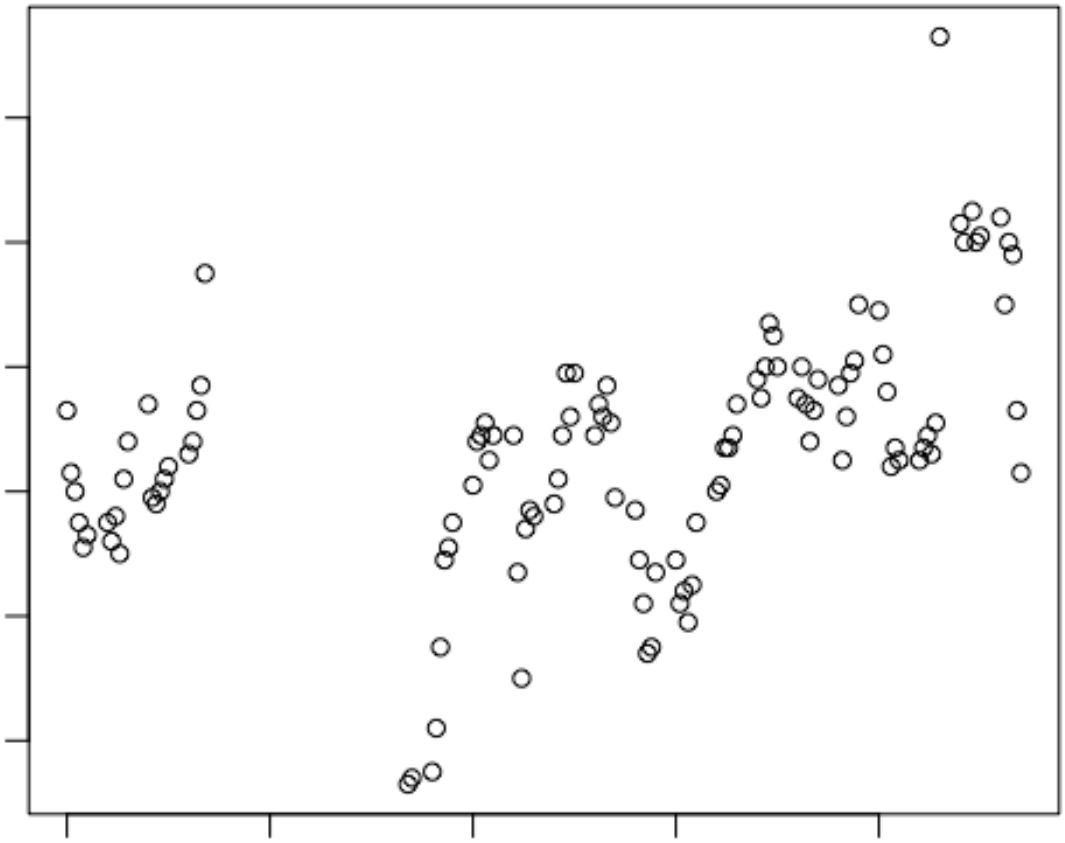
Assignment1

1. The 'beaver1' dataset is included in the default R package.

(a) Make a scatter plot (time on x and temperature on y) of all observations

```
par(mar=c(1, 1, 1, 1))
plot(beaver1$time,beaver1$temp)
```



(b) Confine the data set so that only observations between 8am and 11:59pm are included.

```
B1 <- beaver1[(beaver1$time > 800 & beaver1$time < 1159),]
```

(c) Regress y on x by using the 'lm' function for the dataset obtained in (b). Obtain a summary of the regression.

```
Model.B1 <- lm(B1$temp~B1$time)
summary(Model.B1)

##
## Call:
## lm(formula = B1$temp ~ B1$time)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32543 -0.14337  0.03127  0.13423  0.19642
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.543e+01  3.834e-01  92.411  < 2e-16 ***
## B1$time     1.243e-03  3.791e-04   3.279  0.00417 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1624 on 18 degrees of freedom
## Multiple R-squared:  0.3739, Adjusted R-squared:  0.3391
## F-statistic: 10.75 on 1 and 18 DF,  p-value: 0.004174
```

(d) Briefly discuss what model can formulate the relationship between x and y better than (c).

If we add and apply log to time, the relationship improves seen in the R Squared Value substantially. The model fits the data much better.

```
Model.B1.Improved <- lm(B1$temp~B1$time+log(B1$time))
summary(Model.B1.Improved)
```

```
##
## Call:
## lm(formula = B1$temp ~ B1$time + log(B1$time))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25822 -0.01628  0.03742  0.06453  0.11019
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.295e+02  3.414e+01  -3.792 0.001455 **
## B1$time      -2.679e-02  5.809e-03  -4.612 0.000249 ***
## log(B1$time)  2.795e+01  5.786e+00   4.831 0.000156 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1085 on 17 degrees of freedom
## Multiple R-squared:  0.7361, Adjusted R-squared:  0.7051
## F-statistic: 23.71 on 2 and 17 DF,  p-value: 1.208e-05
```

2. Let(xi,yi)=(0,0),(0.5,0.25),(1,1),(1.5,2.25),(2,4),(2.5,6.25),(3,9).

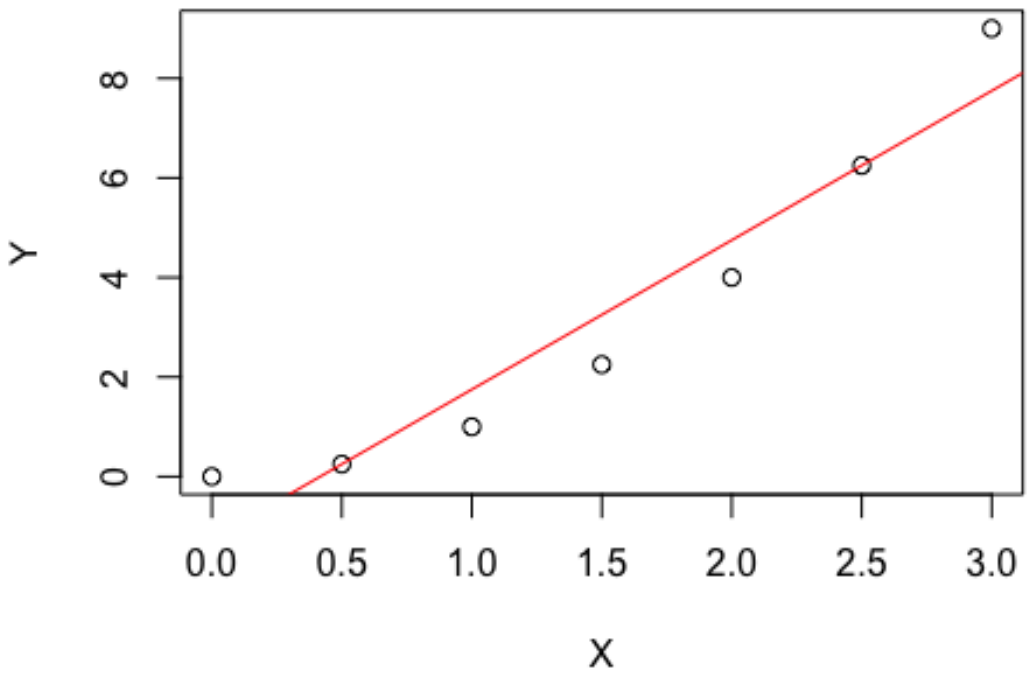
(a) Regress y on x by using the 'lm' function. Obtain a summary of the regression.

```
X <- c(0,.5,1,1.5,2,2.5,3)
Y <- c(0,.25,1,2.25,4,6.25,9)
Q2 <- data.frame(X,Y)
```

```
Model.Q2 <- lm(Q2$Y~Q2$X)
summary(Model.Q2)

##
## Call:
## lm(formula = Q2$Y ~ Q2$X)
##
## Residuals:
##          1          2          3          4          5          6
7
##  1.250e+00 -7.097e-16 -7.500e-01 -1.000e+00 -7.500e-01  2.895e-16  1.250e+
00
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.2500     0.6982  -1.790 0.133416
## Q2$X          3.0000     0.3873   7.746 0.000573 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.025 on 5 degrees of freedom
## Multiple R-squared:  0.9231, Adjusted R-squared:  0.9077
## F-statistic:    60 on 1 and 5 DF,  p-value: 0.0005732
```

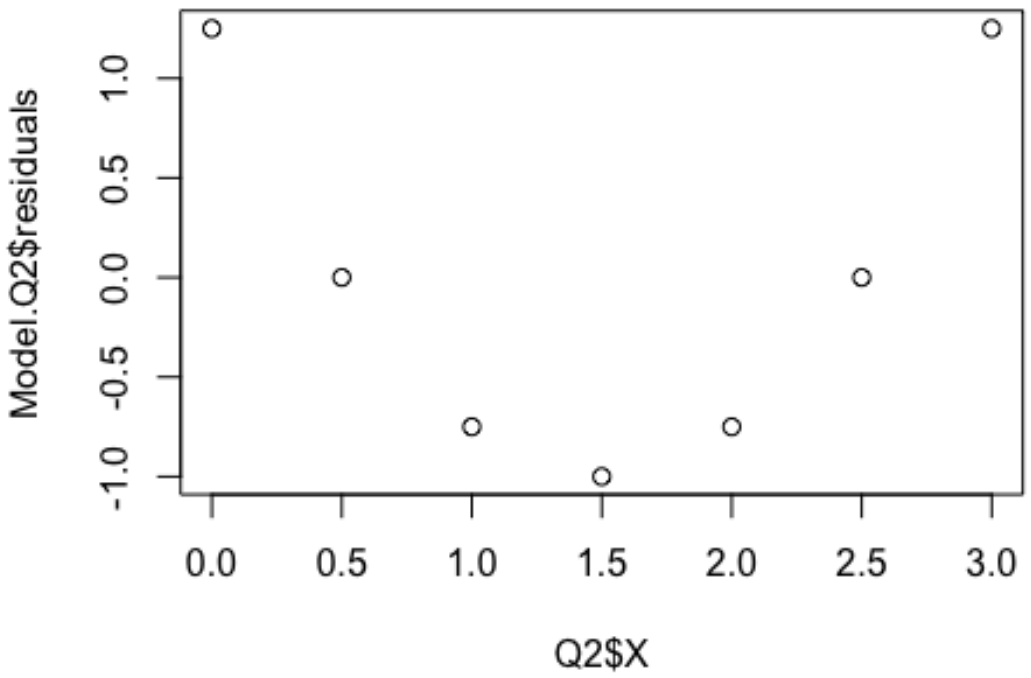
(b) Make a scatter plot with the regression line.

```
par()

plot(Q2)
abline(Model.Q2, col="red")
```



(c) Make a residual plot (the horizontal axis for x, and vertical axis for residuals).

```
plot(Q2$X, Model.Q2$residuals)
```



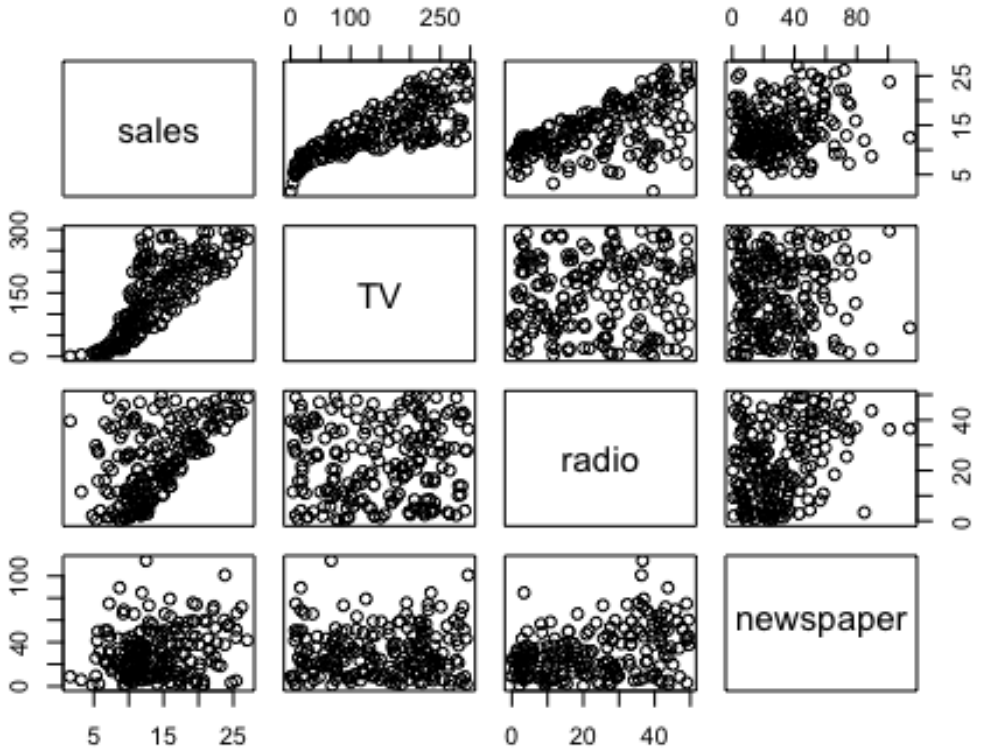
(d) Do you see any problems with the residual plot?

Yes, the residuals are in a quadratic pattern when there should be no pattern.

3. Download and read the data set 'Advertising.csv' into R. We will replicate and explore more details on relationship between sales and advertising on TV shown in Chapter 2.1.

(a) Make a pairwise scatter plot of sales, TV, radio and newspaper by 'pairs' fucntion.

```
Adv <- read.csv("Advertising.csv")
pairs(Adv[,c(5,2,3,4)])
```



(b) Fit a simple linear regression model sales = $\alpha + \beta T\ V + \varepsilon$. Is β significantly different from zero at a 5% significant level?

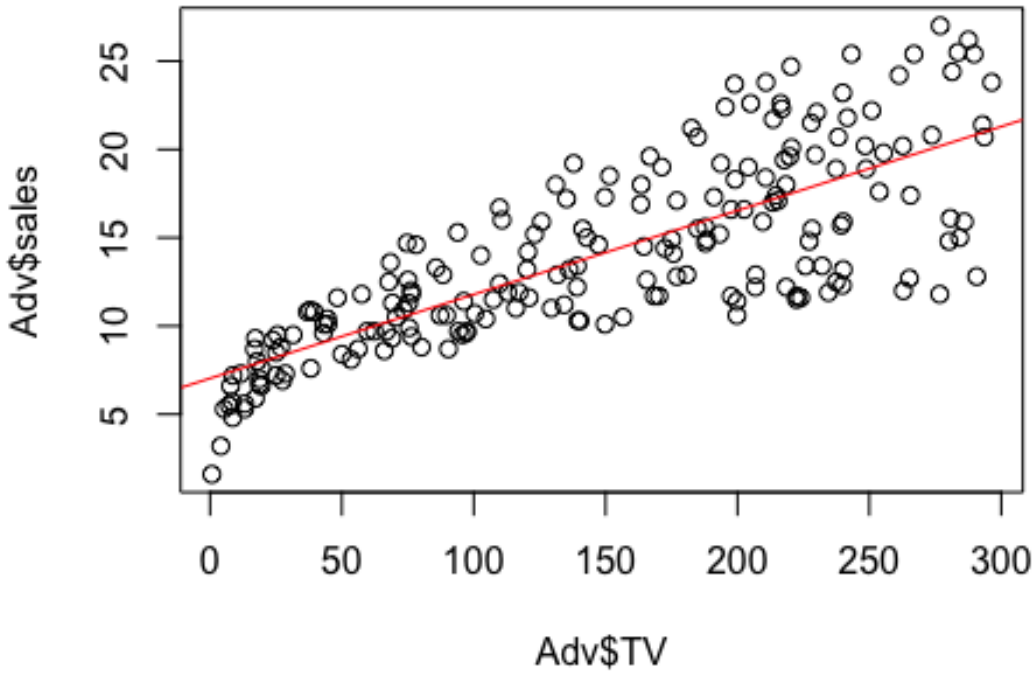
Yes. P-Value is less than .05 so we reject null hypothesis. There is a significant relationship.

```
Model.Adv <- lm(Adv$sales~Adv$TV)
summary(Model.Adv)

##
## Call:
## lm(formula = Adv$sales ~ Adv$TV)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.032594   0.457843   15.36   <2e-16 ***
## Adv$TV      0.047537   0.002691   17.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
```

```
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

(c) Make a scatter plot (for (TV,sales)), and draw a fitted line over the scatter plot in (a).

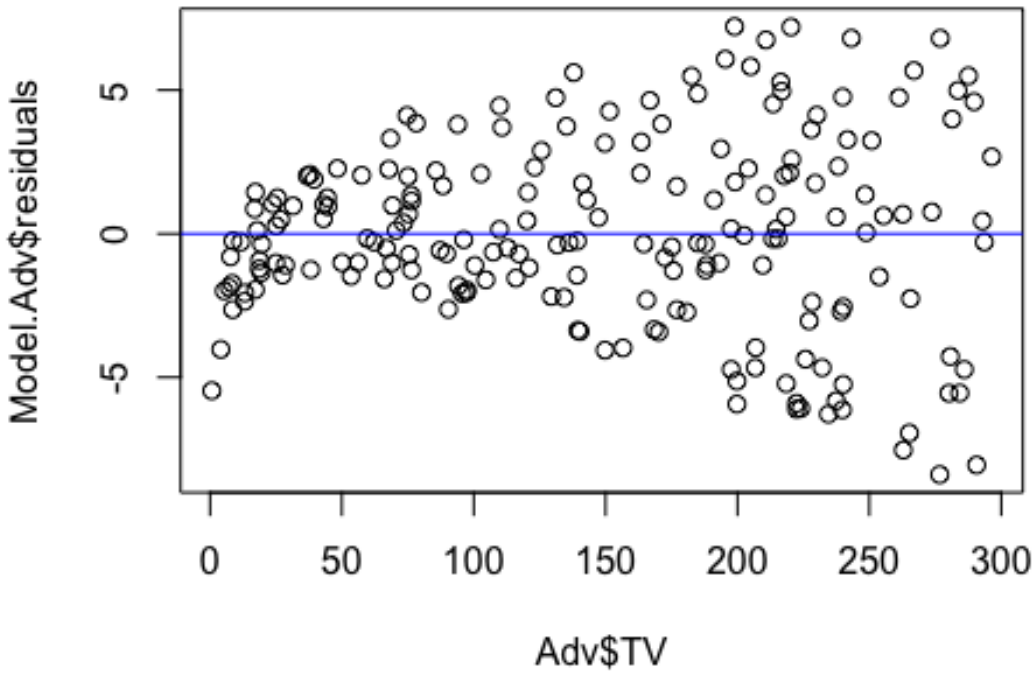
```
plot(Adv$TV, Adv$sales)
abline(Model.Adv, col="red")
```



(d) Make a residual plot (horizontal: TV, vertical: residual), and add the x-axis on it. Comment on whether or not the residuals satisfy the IID assumption of errors.

This does not satisfy IID.The variance is becoming greater over time.

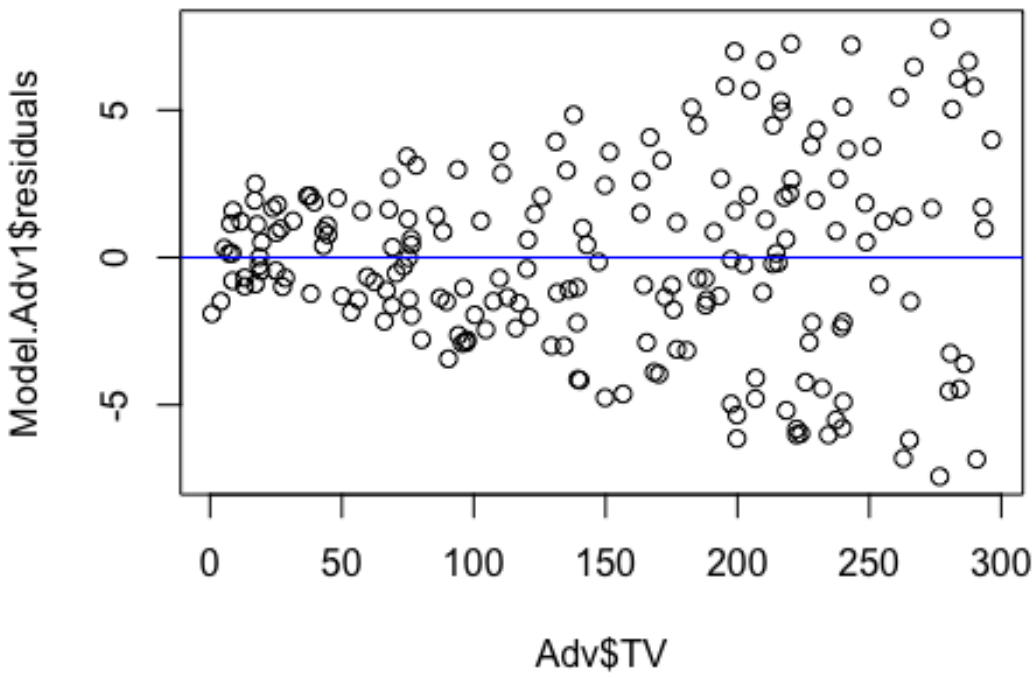
```
plot(Adv$TV,Model.Adv$residuals)
abline(h=0, col="blue")
```



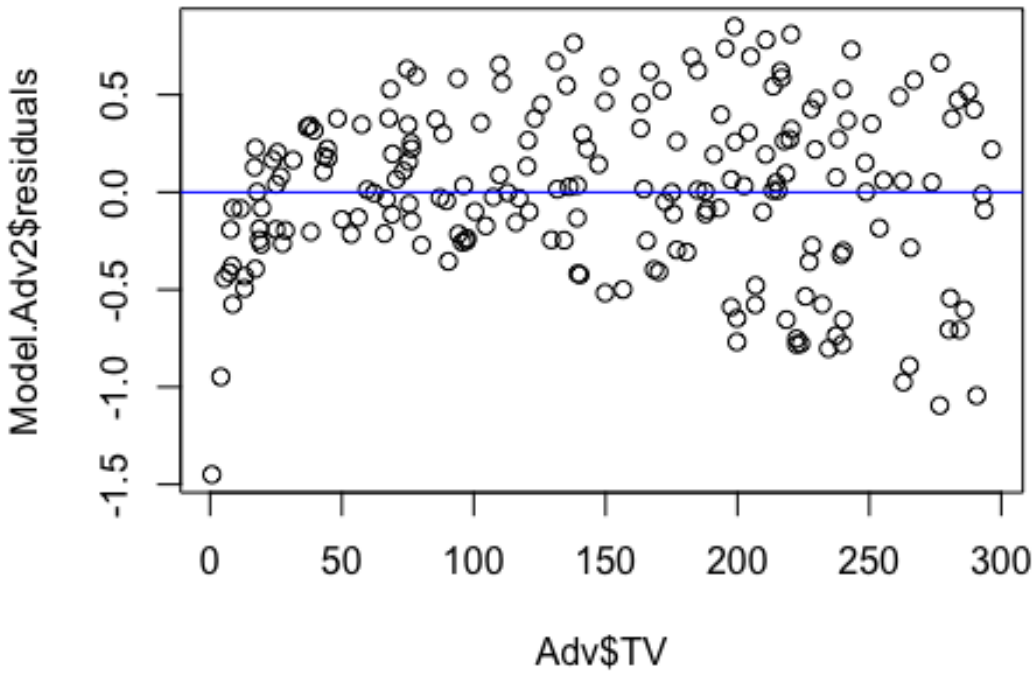
(e) Consider 3 transformations of data. Transform (i) TV by square root, (ii) sales by square root, (iii) both TV and sales by square root. Comment on which scatter plot is relatively better in terms of IID assumption of errors.

The third and final scatter plot is better in terms of IID.

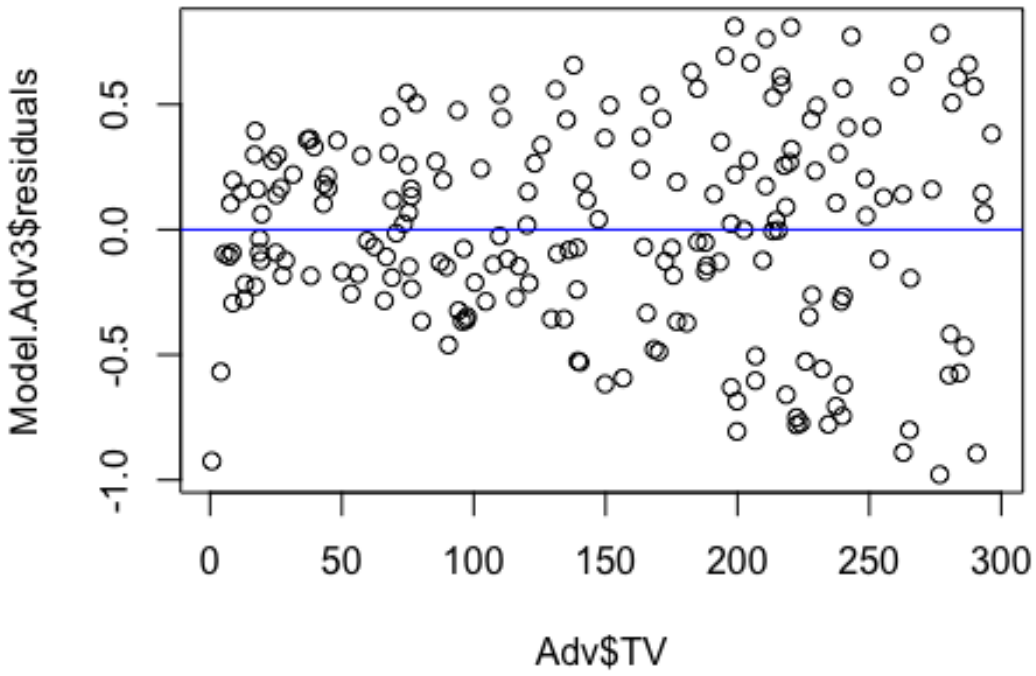
```
Model.Adv1 <- lm(Adv$sales~sqrt(Adv$TV))
Model.Adv2 <- lm(sqrt(Adv$sales)~Adv$TV)
Model.Adv3 <- lm(sqrt(Adv$sales)~sqrt(Adv$TV))

plot(Adv$TV,Model.Adv1$residuals)
abline(h=0, col="blue")
```




```
plot(Adv$TV,Model.Adv2$residuals)
abline(h=0, col="blue")
```



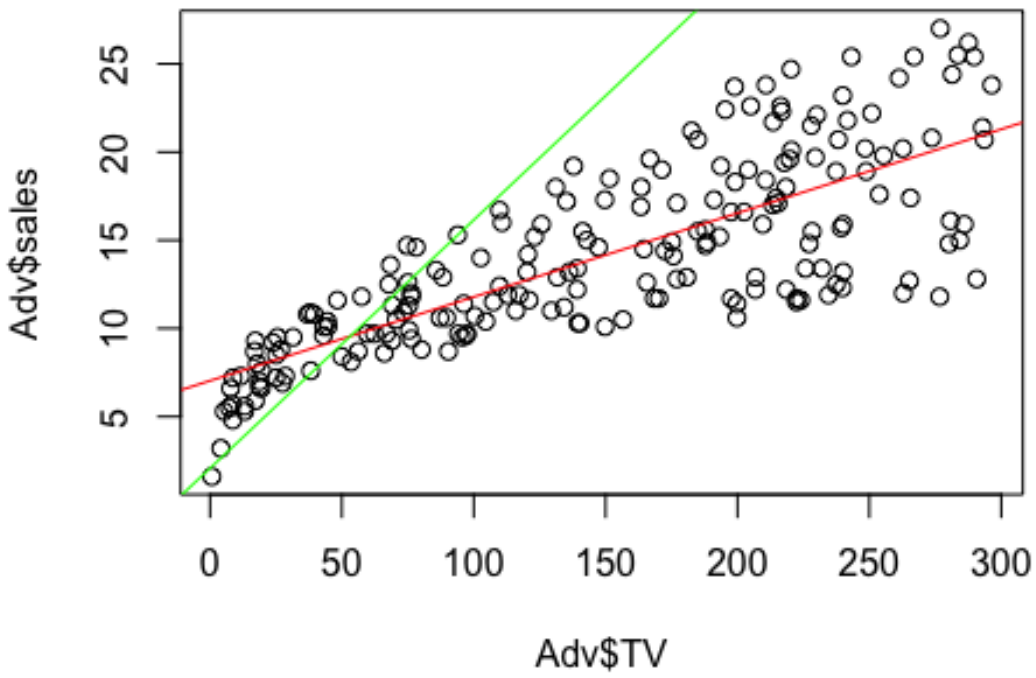

```
plot(Adv$TV,Model.Adv3$residuals)
abline(h=0, col="blue")
```



(f) Implement a simple linear regression again with the transformation you chose in the previous question. Explain if the meaning of the slope changed from (b).

The slope change captures more of the values than the last regression.

```
plot(Adv$TV, Adv$sales)
abline(Model.Adv, col="red")
abline(Model.Adv3, col="green")
```


Adv$sales
5
10
15
20
25
0
50
100
150
200
250
300
Adv$TV