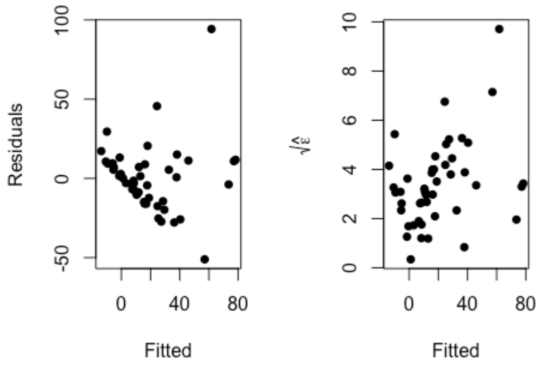
HW#5, Nan Deng

(1)

For normal model:

(a)

```r
library(faraway)
data(teengamb)
attach(teengamb)
teengamb_fit <- lm(gamble~sex+status+income+verbal)
par(mfrow=c(1,2))
plot(fitted(teengamb_fit),resid(teengamb_fit), xlab="Fitted",
     ylab="Residuals",pch=16)
plot(fitted(teengamb_fit), sqrt(abs(resid(teengamb_fit))) ,xlab="Fitted",
     pch=16, ylab=expression(sqrt(hat(epsilon))))
```

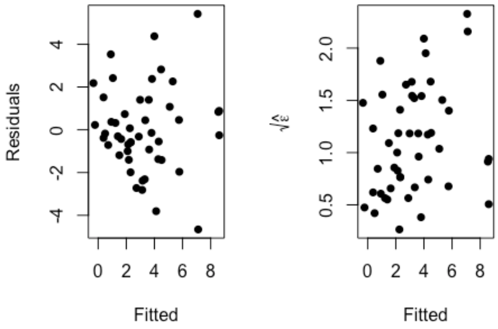


According to the residual distribution against y, it shows Heteroscedasticity (non-constant variance). Although the variance keeps concentrated at the head, it starts to become seperate when x gets greater. Hence, this model does not fit these data.

For square root model:

(a)

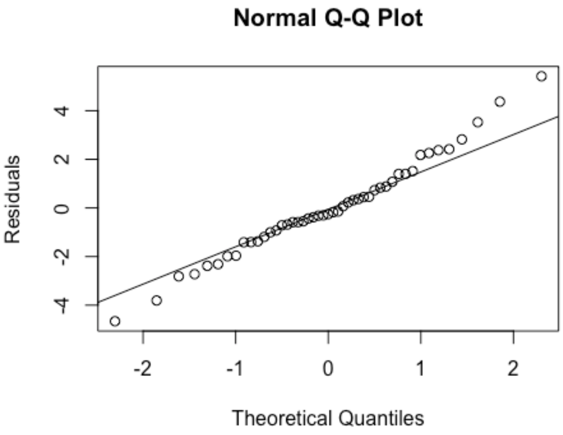
```r
teengamb_sqrt <- lm(sqrt(gamble)~ . , data = teengamb)
par(mfrow=c(1,2))
plot(fitted(teengamb_sqrt),resid(teengamb_sqrt), xlab="Fitted",
     ylab="Residuals",pch=16)
plot(fitted(teengamb_sqrt), sqrt(abs(resid(teengamb_sqrt))) ,xlab="Fitted",
     pch=16, ylab=expression(sqrt(hat(epsilon))))
```



This scale of distribution reveals that the residual tend to show a constant trend towards the model, which mean it is Homoscedasticity.

(b)

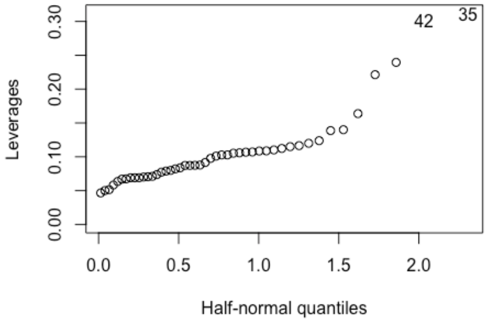
```
qqnorm(resid(teengamb_sqrt), ylab="Residuals")
qqline(resid(teengamb_sqrt))
```



Based on the Normal QQ-plot, this model has a more concentrated tendency according to the scale. Although in this observation some points does not fully fit the estimated line in the head and tail part, they tend to be closer to each other compared to the other model. Therefore, the normal distribution assumption is valid.

(c)

```
halfnorm(hatvalues(teengamb_sqrt), nlab = 2, ylab="Leverages")
```



```
teengamb[c(42,35),]

##    sex status income verbal gamble
## 42   0     61   15.0      9   69.7
## 35   0     28    1.5      1   14.1
```

The large leverage point is 35 and 42.

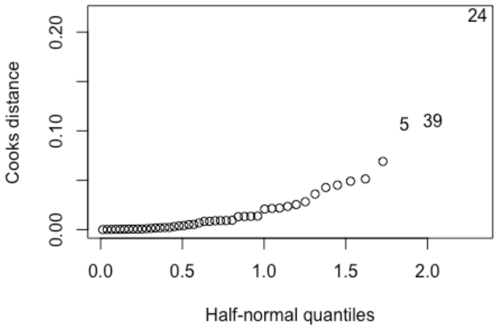
(d)

```
ti_sqrt <- rstudent(teengamb_sqrt)
max(abs(ti_sqrt))

## [1] 3.037005

abs(qt(.05/(47*2),41))

## [1] 3.522795
```

Considering that the maximum distance of points in this model to the line is less than that calculated, this model does not have any outlier.

(e)

```
cook_sqrt <- cooks.distance(teengamb_sqrt)
halfnorm(cook_sqrt, nlab = 3, ylab="Cooks distance")
```

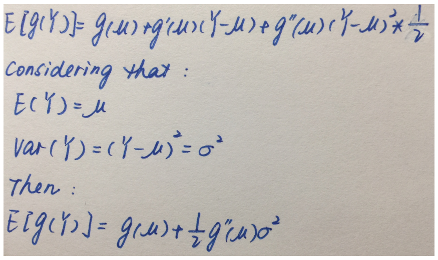


```
cook_sqrt[c(5,24,39)]

##         5        24        39
## 0.1071005 0.2179717 0.1109629
```

The influential point of this model is 5, 24, and 39.

(2)

(a)



$E[g(Y)] = g(\mu) + g'(\mu)(Y-\mu) + g''(\mu)(Y-\mu)^2 * \frac{1}{2}$

Considering that :

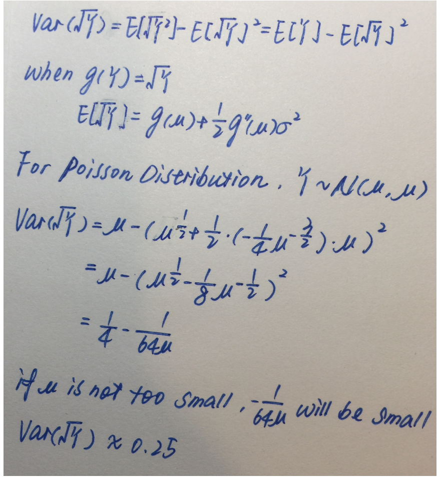
$E(Y) = \mu$

$Var(Y) = (Y-\mu)^2 = \sigma^2$

Then :

$E[g(Y)] = g(\mu) + \frac{1}{2}g''(\mu)\sigma^2$

(b)



$Var(\sqrt{Y}) = E[\sqrt{Y^2}] - E[\sqrt{Y}]^2 = E[Y] - E[\sqrt{Y}]^2$

when $g(Y) = \sqrt{Y}$

$E[\sqrt{Y}] = g(\mu) + \frac{1}{2}g''(\mu)\sigma^2$

For Poisson Disteribution, $Y \sim N(\mu, \mu)$

$$Var(\sqrt{Y}) = \mu - (\mu^{\frac{1}{2}} + \frac{1}{2} \cdot (-\frac{1}{4}\mu^{-\frac{3}{2}}) \cdot \mu)^2$$
$$= \mu - (\mu^{\frac{1}{2}} - \frac{1}{8}\mu^{-\frac{1}{2}})^2$$
$$= \frac{1}{4} - \frac{1}{64\mu}$$

if μ is not too small, $-\frac{1}{64\mu}$ will be small

$Var(\sqrt{Y}) \approx 0.25$