Stat 415/615, Lab 2. A complete Example of Simple Linear Regression

Jun Lu

Stat 415/615 Regression, 2023

Contents

Refer to the Plutonium Measurement case example in Ch.3.11 (p.141). The task here is to estimate the regression relation between alpha count per second (Y) and plutonium activity. The data set can be found on Blackboard (Ch3_11.sav and CH03TA10.txt).

```
pu<-read.table("../DataSets/CH03TA10.txt", header=F)
colnames(pu) <- c("Y.AlphaRate", "X.PuAct")
pu
```

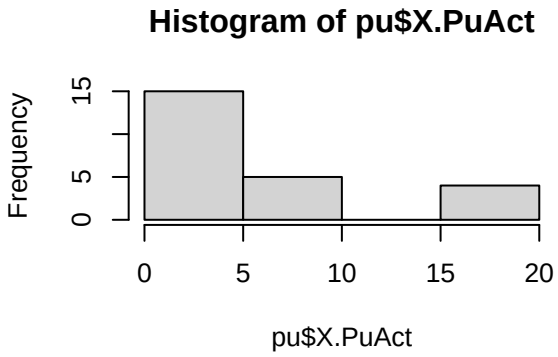
```
##    Y.AlphaRate X.PuAct
## 1        0.150      20
## 2        0.004       0
## 3        0.069      10
## 4        0.030       5
## 5        0.011       0
## 6        0.004       0
## 7        0.041       5
## 8        0.109      20
## 9        0.068      10
## 10       0.009       0
## 11       0.009       0
## 12       0.048      10
## 13       0.006       0
## 14       0.083      20
## 15       0.037       5
```

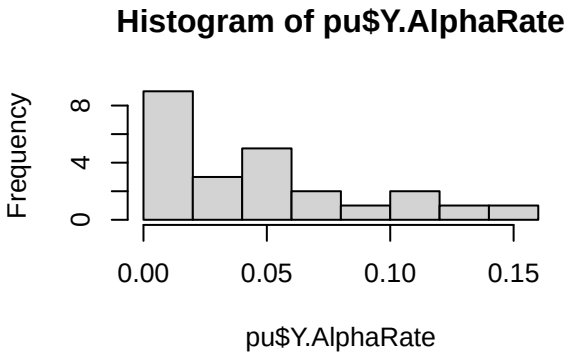
```
## 16       0.039       5
## 17       0.132      20
## 18       0.004       0
## 19       0.006       0
## 20       0.059      10
## 21       0.051      10
## 22       0.002       0
## 23       0.049       5
## 24       0.106       0
```

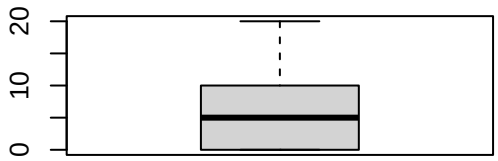
1 Plot data

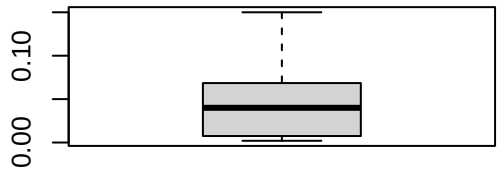
1.1 One variable at a time. (Histogram, box plot, etc.)

```
par(mfrow=c(2,2))
hist(pu$X.PuAct)
hist(pu$Y.AlphaRate)
boxplot(pu$X.PuAct)
boxplot(pu$Y.AlphaRate)
```





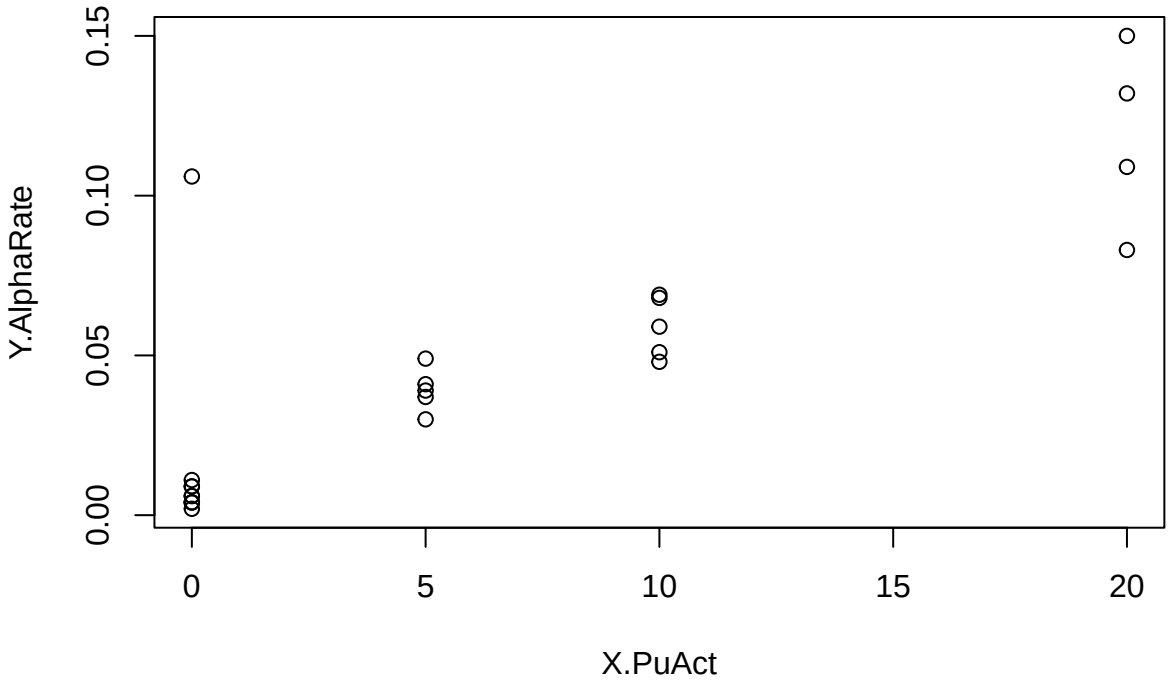




- Both variables are right-skewed. Plutonium activity has a "gap". But it is difficult to claim outliers from the above plots.

1.2 Use scatter plot to check pattern and outliers

```
plot(Y.AlphaRate ~ X.PuAct, data=pu)
```

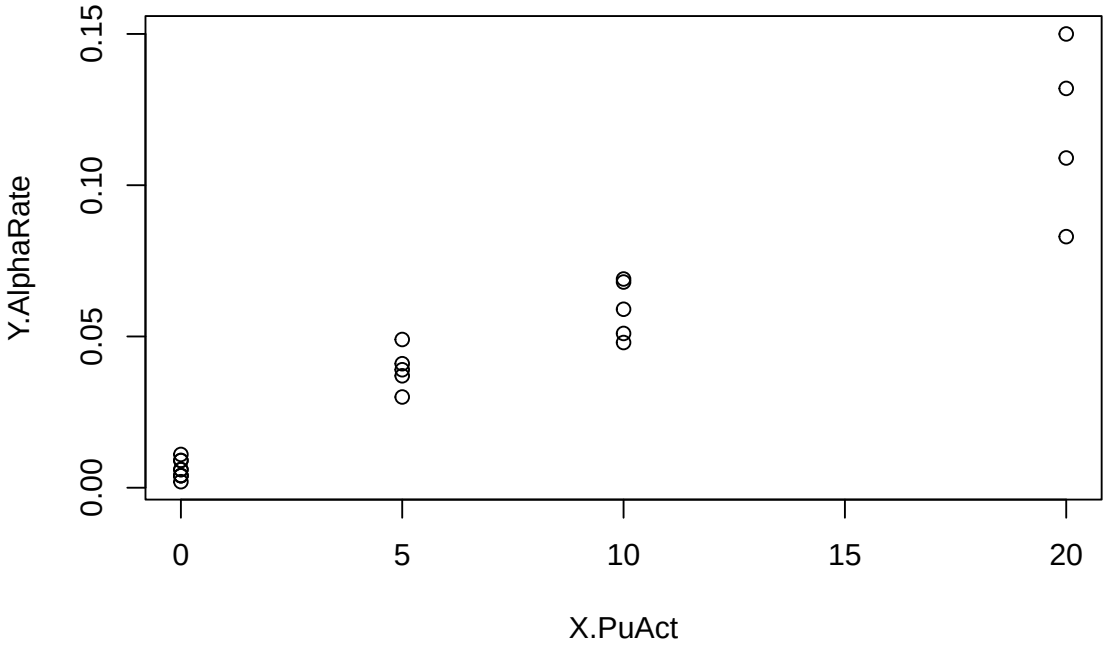


- Findings:
  - There appears to be a positive linear assocation between the variables. The mean of Alpha Rate (Y) increases as Plutonium activity (X) increases.
  - There appears to be a possible outlier ($X = 0$, $Y \approx 0.11$). Note that though the $X = 0$ is within the range of other x-values (0~20), and $Y \approx 0.11$ is within the range of other y-values (0~0.15), ($X = 0$, $Y \approx 0.11$) is clearly deviated from the over-all pattern between X and Y.
  - The variantion of Y gets bigger as X increases.

2 Outlier(s)?

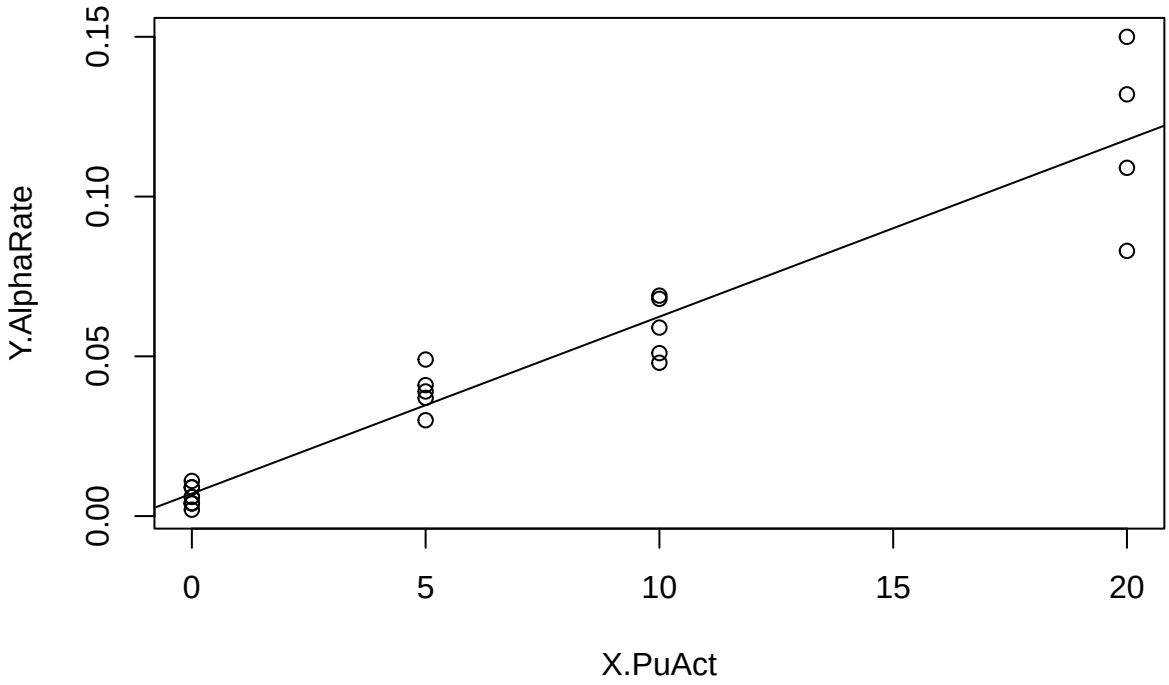
- The above scatter plot shows a possible outlier ($X = 0$, $Y \approx 0.11$). It is observation 24.
- Remove outlier. Then plot again.

```
pu2<-pu[-24, ]
plot(Y.AlphaRate ~ X.PuAct, data=pu2)
```


Y.AlphaRate
0.00
0.05
0.10
0.15
0
5
10
15
20
X.PuAct

3 Fit Regression

```
pu.reg<-lm(Y.AlphaRate ~ X.PuAct, data=pu2)
plot(Y.AlphaRate ~ X.PuAct, data=pu2)
abline(pu.reg$coef)
```

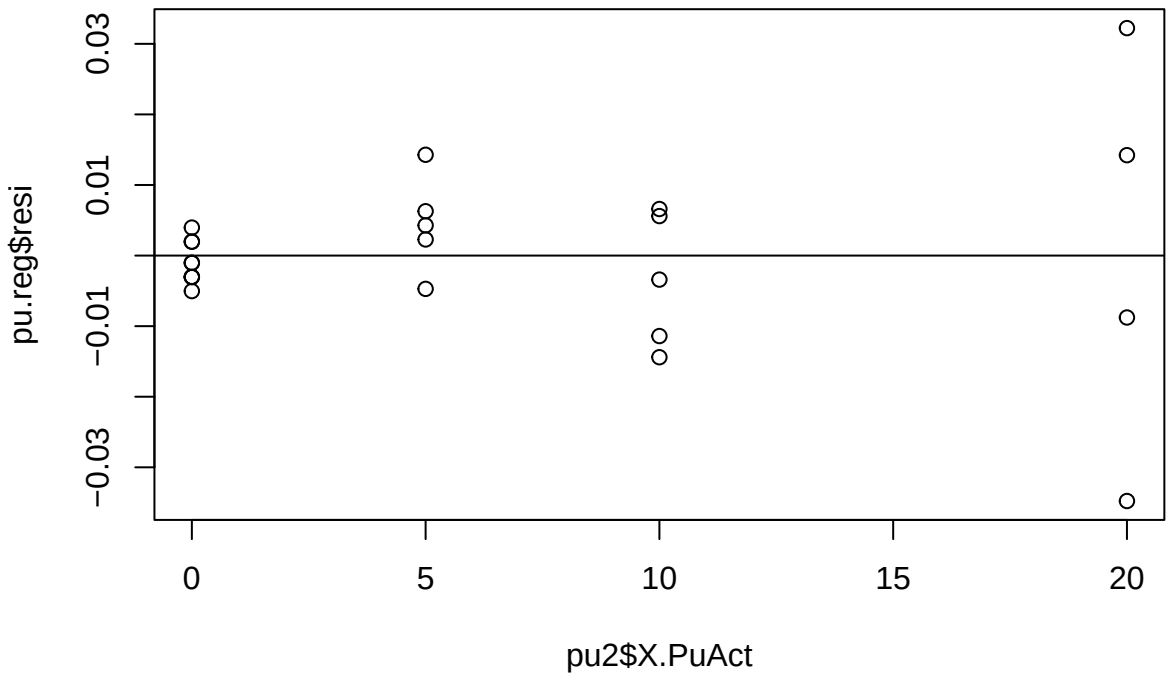


4 Basic diagnosis

4.1 Plot the residuals

4.1.1 The residual plot

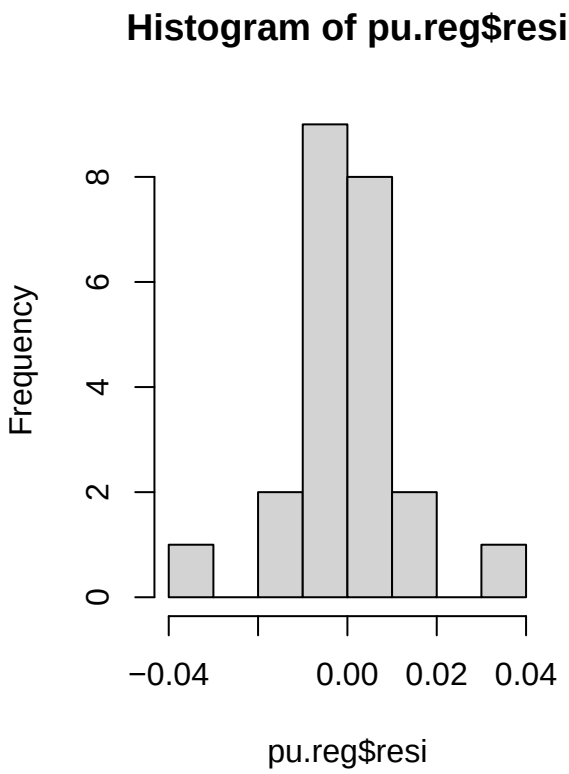
```
plot(pu2$X.PuAct, pu.reg$resi)
abline(0,0)
```

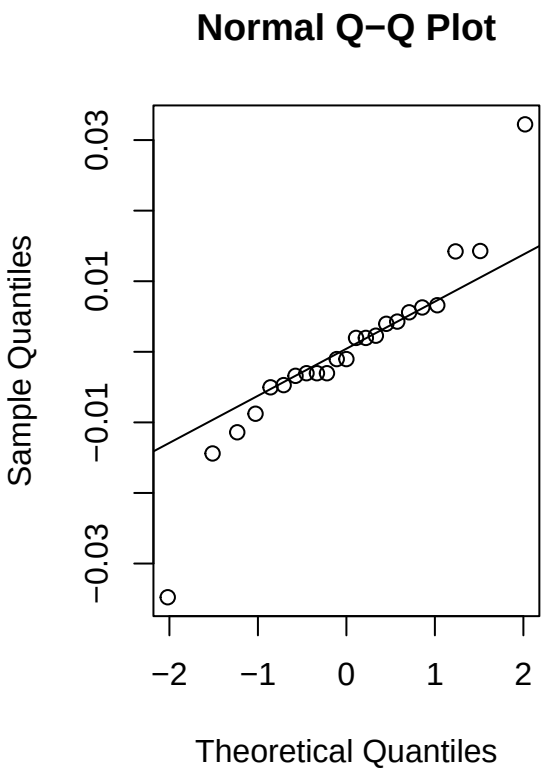


- The residuals scatter around the horizontal line of 0, but they don't form the ideal "horizontal band". The variance of the residuals increases as the predictor (X, Plutonium activity) increases. This is consistent with what we observed earlier in the scatter plot of the data.

4.1.2 Histogram and Normality plot

```
par(mfrow=c(1,2))
hist(pu.reg$resi)
qqnorm(pu.reg$resi)
qqline(pu.reg$resi)
```





- The hisgtogram is bell-shaped and symmetrical. There are gaps on both sides of the histogram, but the gaps are small and don't suggest obvious outliers.
- The dots in the Q-Q plot deviate from the reference line near both lower and uppoer end with some pattern, and are randomly scattered around the reference in the middle section. This suggest the residuals may not follow Normal distribution, especially near the tails.

4.2 Test of Normaility

```
shapiro.test(pu.reg$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  pu.reg$residuals
## W = 0.91273, p-value = 0.04665
```

- At p-value 0.047, we reject the null hypothesis (H_0: residuals follow Normal distribution). Since there is significance evidence to believe the residuals are not Normally distributed, we'll consider transformation.

4.3 Test of linearity (lack of fit)

```
pu.lof<-lm(Y.AlphaRate ~ as.factor(X.PuAct), data=pu2)
anova(pu.reg, pu.lof)
```

```
## Analysis of Variance Table
##
```

```
## Model 1: Y.AlphaRate ~ X.PuAct
## Model 2: Y.AlphaRate ~ as.factor(X.PuAct)
##   Res.Df       RSS Df  Sum of Sq      F Pr(>F)
## 1     21 0.0033188
## 2     19 0.0031507  2 0.00016811 0.5069 0.6103
```

- Verify the expanded ANOVA table for test of linearity.
  - From the output: $SSE = 0.0033188$, $SS_{PE} = 0.0031507$.
  - $SS_{LF} = SSE - SS_{PE} = 0.0033188 - 0.0031507 = 0.00016811$.
  - Sample size $n = 23$, with 4 unique x-values. Hence, $df_E = n - 2 = 23 - 2 = 21$, $df_{PE} = n - c = 23 - 4 = 19$, $df_{LF} = c - 2 = 4 - 2 = 2$. $df_E = df_{LF} + df_{PE}$.
  - The F-statistic for this test is

$$\begin{aligned} F_{LF} &= \frac{MS_{LF}}{MS_{PE}} = \frac{SS_{LF}/df_{LF}}{SS_{PE}/df_{PE}} \\ &= \frac{0.00016811/2}{0.0031507/19} \\ &= 0.5069 \end{aligned}$$

  - p-value = $P(F_{(df_1=df_{LF}=2,\ df_2=df_{PE}=19)} > 0.5069) = 0.6103$.
  - The formula-based calculation results are the same as the results in the output table.
- At p-value 0.61, we do not reject the null hypothesis ($H_0 : E(Y) = \beta_0 + \beta_1 X$, i.e., model fits data).

5 Transform data to improve Model 1, and get Model 2.

Question: Why do we transform Y?

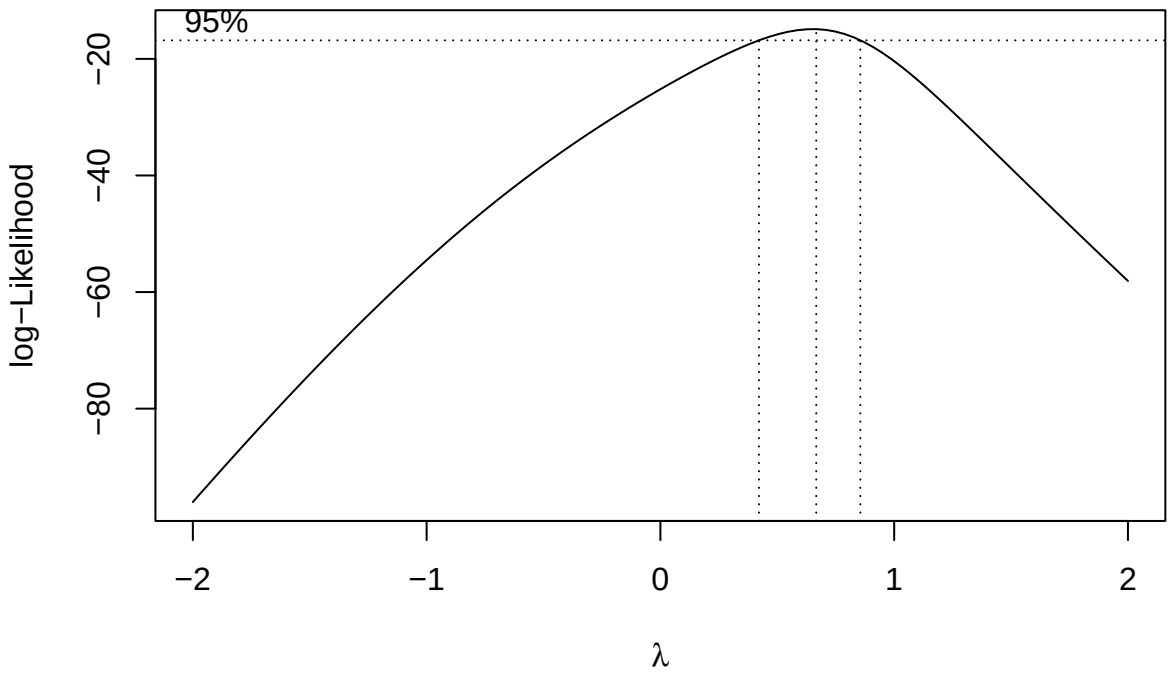
Answer: To address the non-constant variance.

```
pu2$sqrtY<-sqrt(pu2$Y.AlphaRate)
pu.reg2<-lm(sqrtY ~ X.PuAct, data=pu2)
```

- (Optional) Use MLE to determine the box-cox transformation

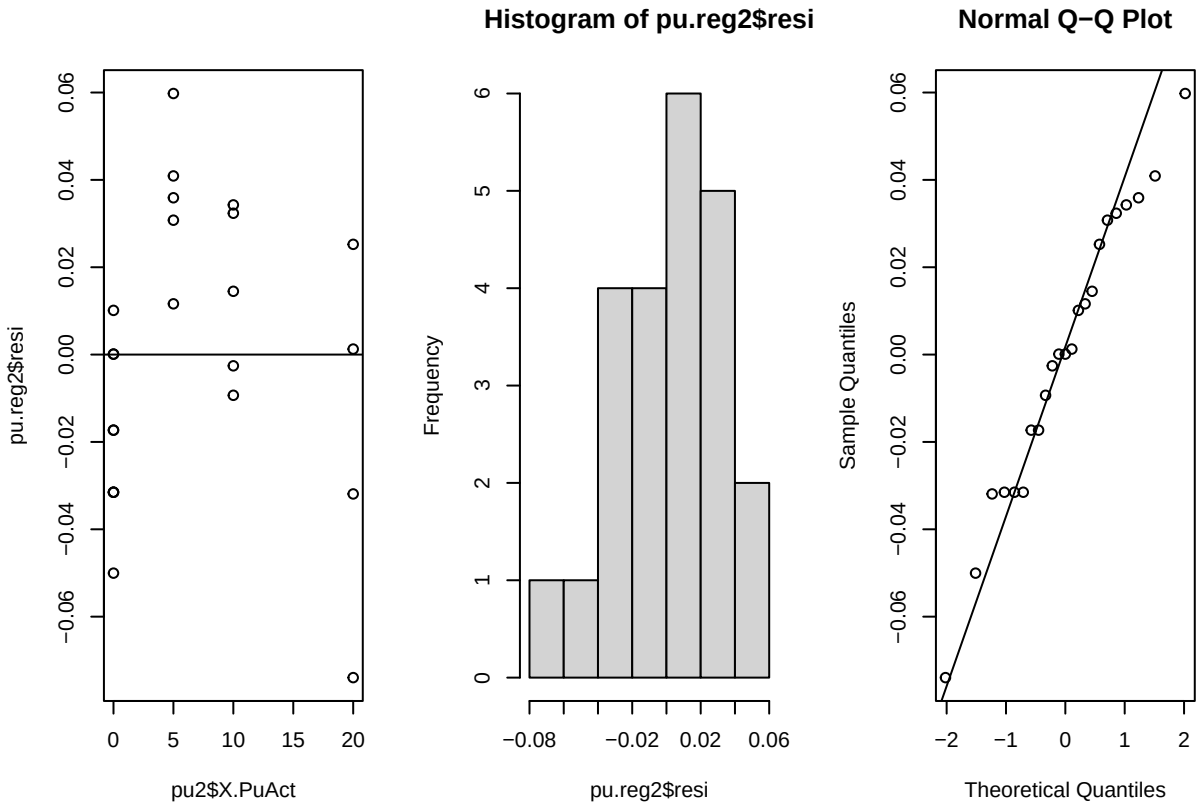
```
# Download and install the package. You only need to do this once.
install.packages("MASS")
```

```
library(MASS)
boxcox(pu.reg) ## Note that 0.5 within the 95% CI.
```



6 Diagnose (i.e., check model assumptions) Model 2

```
par(mfrow=c(1,3))
plot(pu2$X.PuAct, pu.reg2$resi)
abline(0,0)
hist(pu.reg2$resi)
qqnorm(pu.reg2$resi)
qqline(pu.reg2$resi)
```




```
shapiro.test(pu.reg2$resi)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  pu.reg2$resi
## W = 0.97737, p-value = 0.8568
```

```
pu.lof2<-lm(sqrtY ~ as.factor(X.PuAct), data=pu2)
anova(pu.reg2, pu.lof2)  ## Lack of fit test
```

```
## Analysis of Variance Table
##
## Model 1: sqrtY ~ X.PuAct
## Model 2: sqrtY ~ as.factor(X.PuAct)
##   Res.Df      RSS Df Sum of Sq      F  Pr(>F)
## 1     21 0.023453
## 2     19 0.011346  2  0.012106 10.136 0.00101 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

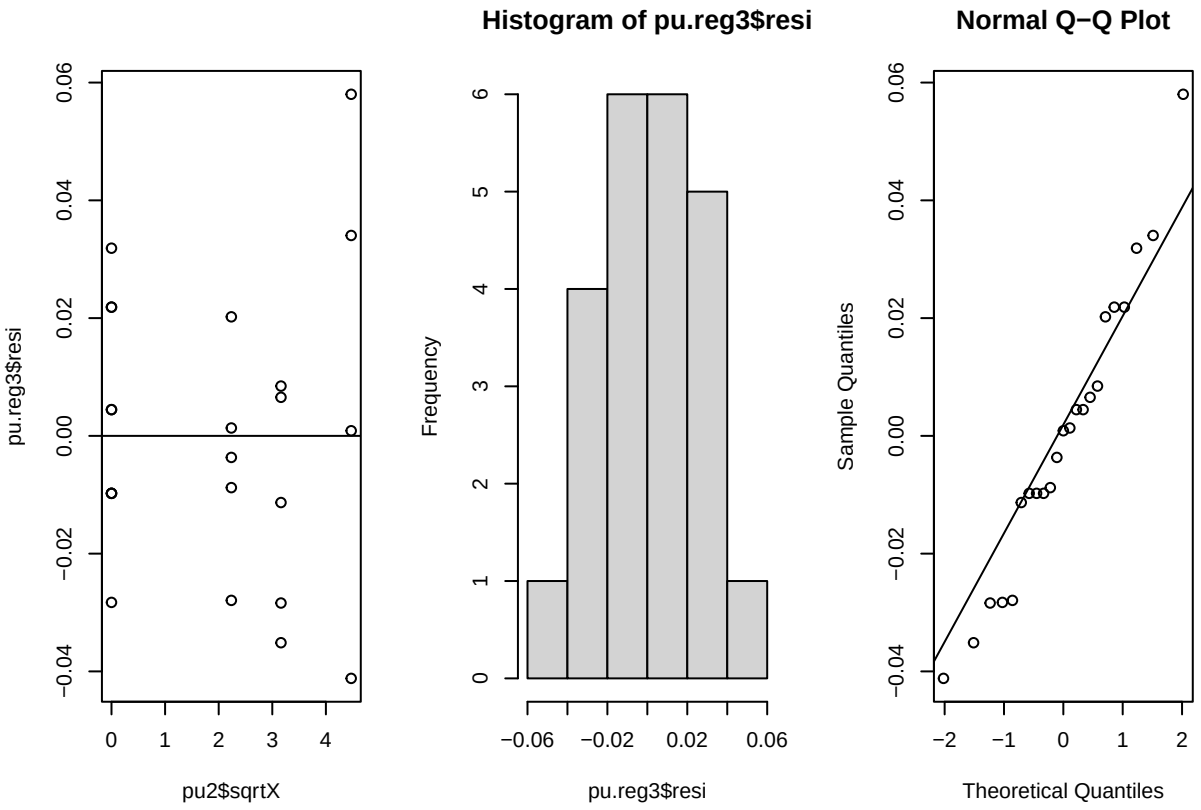
- We see that in the estimated model 2: $\widehat{\sqrt{Y}} = \hat{\beta}_0 + \hat{\beta}_1 X$, the resdidual variance appear to be constant.
- Though the histogram of the residuals appears to be slightly skewed to the left, the Normal Q-Q plot and the Shaprio_Wilk test suggest we do not need to reject the Normality. I.e., we can assume the residuals are Normally distributed.
- Both the residual plot and the test for linearity suggest the linear assumption in model 2 (between $\sqrt{Y}$ and X) is invalid. We will revise the model further.

7 Improve Model 2 to get Model 3. Check the model assumptions on Model 3.

```
pu2$sqrtX<-sqrt(pu2$X.PuAct)
pu.reg3<-lm(sqrtY ~ sqrtX, data=pu2)
summary(pu.reg3)
```

```
##
## Call:
## lm(formula = sqrtY ~ sqrtX, data = pu2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.041186 -0.010541  0.000868  0.014336  0.058015
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.073006   0.007831   9.323 6.51e-09 ***
## sqrtX       0.057305   0.003016  18.998 1.05e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02477 on 21 degrees of freedom
## Multiple R-squared:  0.945,  Adjusted R-squared:  0.9424
## F-statistic: 360.9 on 1 and 21 DF,  p-value: 1.046e-14
```

```
par(mfrow=c(1,3))
plot(pu2$sqrtX, pu.reg3$resi)
abline(0,0)
hist(pu.reg3$resi)
qqnorm(pu.reg3$resi)
qqline(pu.reg3$resi)
```




```
shapiro.test(pu.reg3$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  pu.reg3$residuals
## W = 0.97091, p-value = 0.7109
```

```
pu.lof3<-lm(sqrtY ~ as.factor(sqrt(X.PuAct)), data=pu2)
anova(pu.reg3, pu.lof3)
```

```
## Analysis of Variance Table
##
## Model 1: sqrtY ~ sqrtX
## Model 2: sqrtY ~ as.factor(sqrt(X.PuAct))
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1     21 0.012883
## 2     19 0.011346  2 0.0015368 1.2868 0.2992
```

```
summary(pu.reg3)
```

```
##
## Call:
## lm(formula = sqrtY ~ sqrtX, data = pu2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.041186 -0.010541  0.000868  0.014336  0.058015
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.073006   0.007831   9.323 6.51e-09 ***
## sqrtX       0.057305   0.003016  18.998 1.05e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02477 on 21 degrees of freedom
## Multiple R-squared:  0.945,  Adjusted R-squared:  0.9424
## F-statistic: 360.9 on 1 and 21 DF,  p-value: 1.046e-14
```

```
anova(pu.reg3)
```

```
## Analysis of Variance Table
##
## Response: sqrtY
##           Df   Sum Sq  Mean Sq F value    Pr(>F)
## sqrtX      1 0.221416 0.221416  360.92 1.046e-14 ***
## Residuals 21 0.012883 0.000613
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Are the model assumptions satisfied? **YES!**
- Model 3 passed our check on linearity, constant variance and Normality assumptions. We removed the outlier at the beginning.
- We did not check other predictors because we do not have other variables. We did not check the independence because there is no time variable.

8 What is the final regression model you would recommend?

The final estimated model is: $\widehat{\sqrt{Y}} = 0.073 + 0.057\sqrt{X}$.

Express the estimated regression model in its original scale:

$$\widehat{Y} = (0.073 + 0.057\sqrt{X})^2$$

—- This is the end of Lab 2. —-