Classwork/Lab Week 4 Session 1

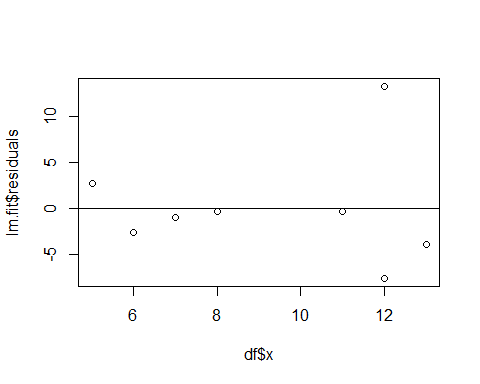
Johnson ODEJIDE

2023-02-07

## 1

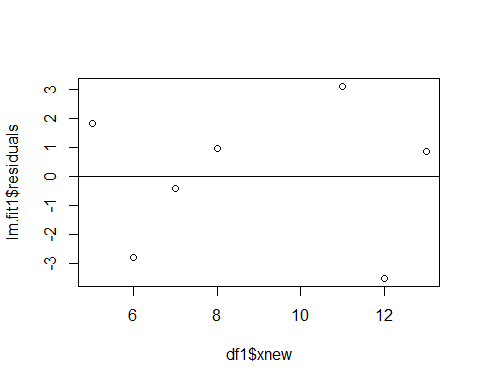
Plot the residuals against ei against xi

x <- c(5, 8, 11, 7, 13, 12, 12, 6)  
  
y <- c(63, 67, 74, 64, 75, 69, 90, 60)  
  
df <- tibble(x, y)  
  
lm.fit <- lm(y ~ x)  
  
plot(df$x, lm.fit$residuals)  
abline(0, 0)

 The data does not appear to be normal because the variance are not constant, most of the data points are below the line.

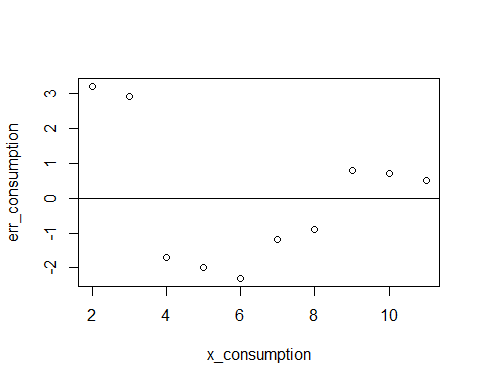
### b

xnew <- x[-7]  
ynew <- y[-7]  
df1 <- tibble(xnew, ynew)  
  
lm.fit1 <- lm(ynew ~ xnew)  
plot(df1$xnew, lm.fit1$residuals)  
abline(0, 0)

 The data appears to be more normal than the previous model. There appears to be an equal distribution of the data points below and above the line. Although, a closer look at the points indicate a form of curvature. In order to conclude, other plots might be plotted.

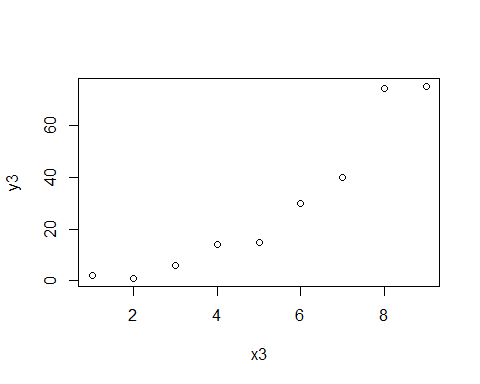
## 2

x\_consumption <- c(2:11)  
err\_consumption <- c(3.2, 2.9, -1.7, -2.0, -2.3, -1.2, -0.9, 0.8, 0.7, 0.5)  
  
plot(x\_consumption, err\_consumption)  
abline(0, 0)

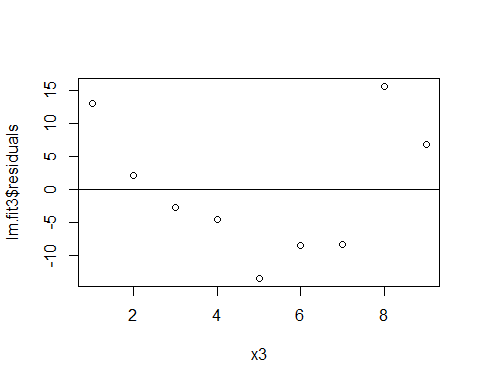
 Yes, a transformation might alleviate this problem.

## 3

x3 <- c(1:9)  
y3 <- c(2, 1, 6, 14, 15, 30, 40, 74, 75)  
  
lm.fit3 <- lm(y3 ~ x3)  
plot(x3, y3)



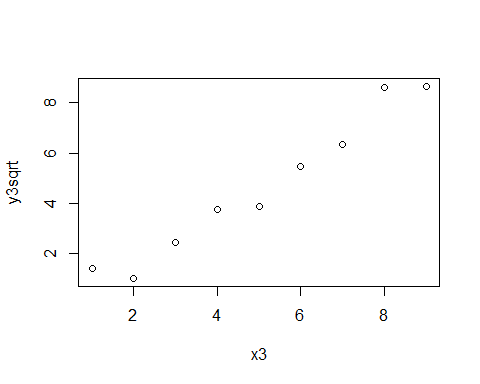
plot(x3, lm.fit3$residuals)  
abline(0, 0)



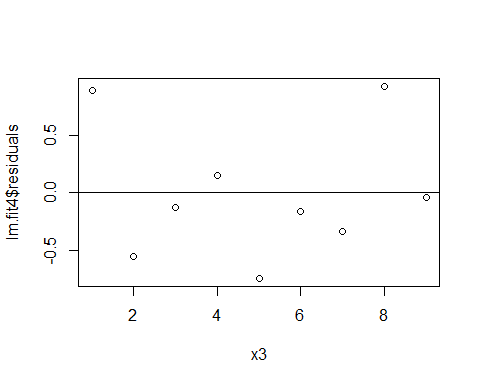
A linear model is not appropriate because there seems to be a curvature downward up in a positive trend on the scatter plot, showing that the data is skewed right. Furthermore, in the residual plot, the variance does not appear to be constant, most of the points are below the line.

## 3b

y3sqrt <- sqrt(y3)  
plot(x3, y3sqrt)



lm.fit4 <- lm(y3sqrt ~ x3)  
# summary(lm.fit4)  
plot(x3, lm.fit4$residuals)  
abline(0, 0)

 A look at the scatter plot indicates that a linear regression model is more appropriate with the transformed data than the previous one although not perfect, there are more points that are closer to the regression line. Furthermore, the residual plot also indicates some form that generally shows the data is more appropriate given the number of data points available in the data.

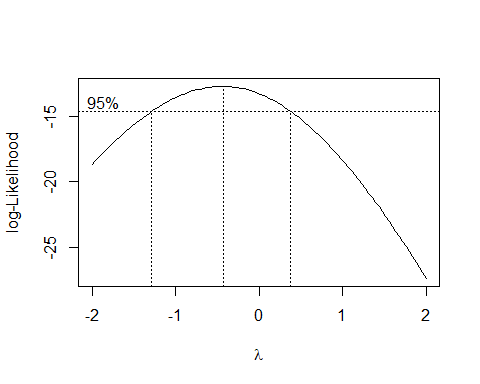
## 4

For the bivariate data given below, the residual plots suggest problems involving non-normality or non-constant error variance or both. Use Box – Cox method as indicated in class to produce a lamda power transformation that will best normalize the data. Your work should include all graphs and plots that support your final answer. Use a series of steps and reasoning demonstrated in class.

x4 <- c(7, 7, 8, 3, 2, 4, 4, 6, 6, 7, 5, 3, 3, 5, 8)  
y4 <- c(1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 6, 7, 8)  
  
boxcox <- data.frame(x4, y4)  
  
model <- lm(y4 ~ x4)   
model

##   
## Call:  
## lm(formula = y4 ~ x4)  
##   
## Coefficients:  
## (Intercept) x4   
## 3.000e+00 2.709e-16

library(MASS)  
box\_cox <- boxcox(y4 ~ x4)



box\_cox

## $x  
## [1] -2.00000000 -1.95959596 -1.91919192 -1.87878788 -1.83838384 -1.79797980  
## [7] -1.75757576 -1.71717172 -1.67676768 -1.63636364 -1.59595960 -1.55555556  
## [13] -1.51515152 -1.47474747 -1.43434343 -1.39393939 -1.35353535 -1.31313131  
## [19] -1.27272727 -1.23232323 -1.19191919 -1.15151515 -1.11111111 -1.07070707  
## [25] -1.03030303 -0.98989899 -0.94949495 -0.90909091 -0.86868687 -0.82828283  
## [31] -0.78787879 -0.74747475 -0.70707071 -0.66666667 -0.62626263 -0.58585859  
## [37] -0.54545455 -0.50505051 -0.46464646 -0.42424242 -0.38383838 -0.34343434  
## [43] -0.30303030 -0.26262626 -0.22222222 -0.18181818 -0.14141414 -0.10101010  
## [49] -0.06060606 -0.02020202 0.02020202 0.06060606 0.10101010 0.14141414  
## [55] 0.18181818 0.22222222 0.26262626 0.30303030 0.34343434 0.38383838  
## [61] 0.42424242 0.46464646 0.50505051 0.54545455 0.58585859 0.62626263  
## [67] 0.66666667 0.70707071 0.74747475 0.78787879 0.82828283 0.86868687  
## [73] 0.90909091 0.94949495 0.98989899 1.03030303 1.07070707 1.11111111  
## [79] 1.15151515 1.19191919 1.23232323 1.27272727 1.31313131 1.35353535  
## [85] 1.39393939 1.43434343 1.47474747 1.51515152 1.55555556 1.59595960  
## [91] 1.63636364 1.67676768 1.71717172 1.75757576 1.79797980 1.83838384  
## [97] 1.87878788 1.91919192 1.95959596 2.00000000  
##   
## $y  
## [1] -18.65840 -18.38718 -18.12041 -17.85823 -17.60076 -17.34814 -17.10051  
## [8] -16.85800 -16.62075 -16.38893 -16.16266 -15.94210 -15.72741 -15.51873  
## [15] -15.31623 -15.12006 -14.93038 -14.74735 -14.57113 -14.40187 -14.23974  
## [22] -14.08490 -13.93749 -13.79768 -13.66562 -13.54145 -13.42532 -13.31737  
## [29] -13.21774 -13.12656 -13.04395 -12.97004 -12.90493 -12.84873 -12.80154  
## [36] -12.76344 -12.73452 -12.71485 -12.70448 -12.70347 -12.71187 -12.72969  
## [43] -12.75698 -12.79373 -12.83995 -12.89562 -12.96075 -13.03528 -13.11919  
## [50] -13.21243 -13.31495 -13.42667 -13.54753 -13.67744 -13.81633 -13.96408  
## [57] -14.12062 -14.28581 -14.45957 -14.64177 -14.83229 -15.03100 -15.23778  
## [64] -15.45251 -15.67504 -15.90525 -16.14300 -16.38816 -16.64058 -16.90014  
## [71] -17.16669 -17.44011 -17.72025 -18.00699 -18.30018 -18.59971 -18.90544  
## [78] -19.21724 -19.53499 -19.85857 -20.18785 -20.52272 -20.86306 -21.20875  
## [85] -21.55968 -21.91575 -22.27685 -22.64287 -23.01372 -23.38928 -23.76947  
## [92] -24.15420 -24.54336 -24.93687 -25.33465 -25.73660 -26.14264 -26.55271  
## [99] -26.96670 -27.38455

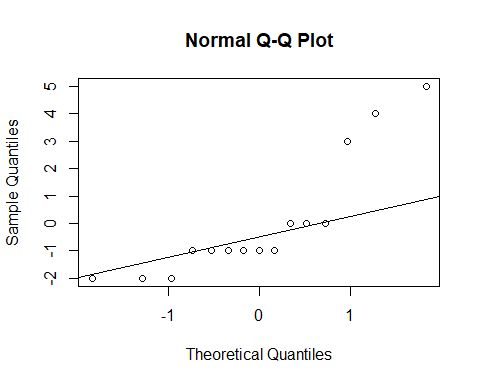
lambda <- box\_cox$x[which.max(box\_cox$y)]   
lambda

## [1] -0.4242424

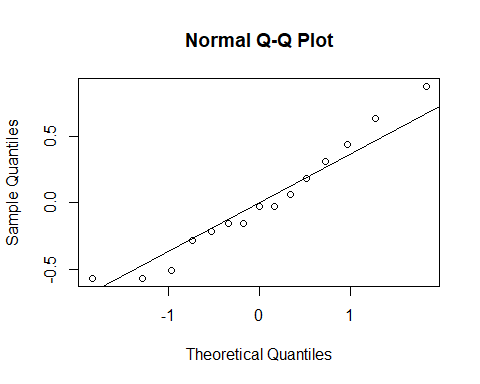
new\_model <- lm(((y4^lambda-1)/lambda) ~ x4)   
new\_model

##   
## Call:  
## lm(formula = ((y4^lambda - 1)/lambda) ~ x4)  
##   
## Coefficients:  
## (Intercept) x4   
## 1.00564 -0.06264

qqnorm(model$residuals)   
qqline(model$residuals)



qqnorm(new\_model$residuals)   
qqline(new\_model$residuals)



The highest lamda power transformation is ***-0.4242424***