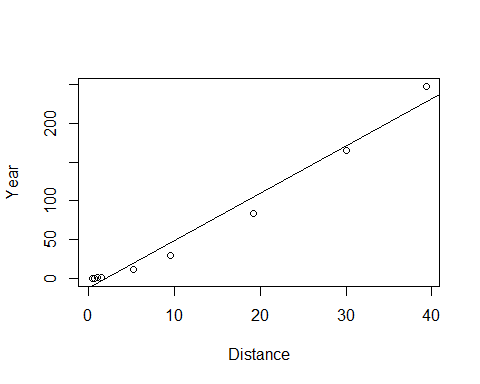
STOR 455 Class 4 assessing conditions and transformations

# message=FALSE, warning=FALSE suppress warnings and messages from appearing in knitted html  
  
library(readr)  
library(Stat2Data)  
  
Planets <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data//Planets.csv")

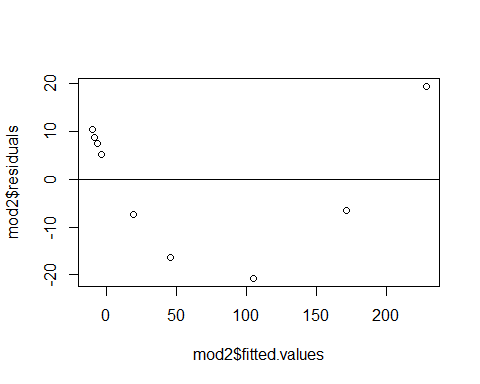
head(Planets, 9)

## # A tibble: 9 x 7  
## Planet Distance Year Mass Day Diameter Gravity  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Mercury 0.39 0.24 0.055 1408. 3.04 0.37  
## 2 Venus 0.72 0.61 0.815 5832. 7.52 0.88  
## 3 Earth 1 1 1 24 7.92 1   
## 4 Mars 1.52 1.88 0.107 24.6 4.22 0.17  
## 5 Jupiter 5.2 11.9 318. 9.9 88.8 2.64  
## 6 Saturn 9.52 29.5 95.2 10.2 74.6 1.15  
## 7 Uranus 19.2 84.1 14.5 17.2 31.6 1.15  
## 8 Neptune 30.0 165. 17.2 16.1 30.2 1.12  
## 9 Pluto 39.3 248. 0.003 153. 1.86 0.04

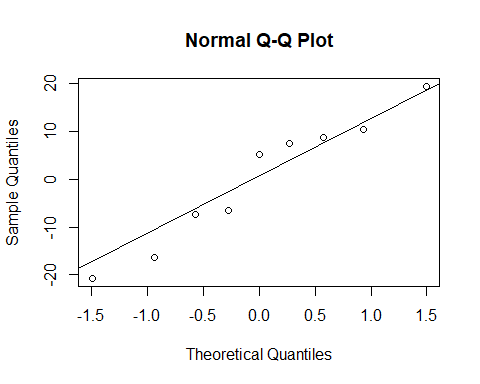
plot(Year~Distance, data=Planets)  
mod2=lm(Year~Distance, data=Planets)  
abline(mod2)

 *Simple Linear Model- Conditions* **Model:** 1. Linearity: The means for Y vary as a linear function of X. **Error:** 2.Zero Mean: The distribution of the errors is centered at zero. 3.Constant variance: The variance for Y is the same at each X. (Homoscedasticity) 4.Independence: No relationships among errors. 5.Normality: - Residuals are normally distributed - (sometimes) At each X, the Y’s follow a normal distribution.

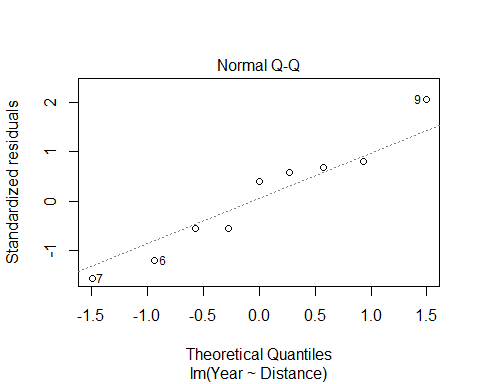
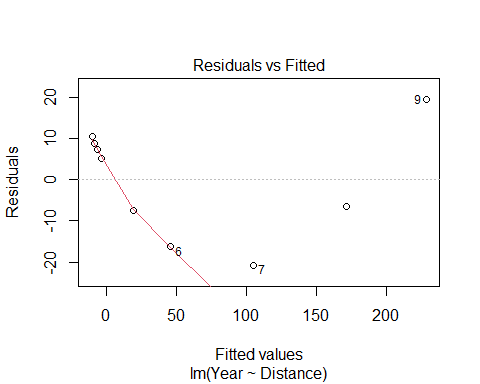
plot(mod2$residuals~mod2$fitted.values)  
abline(0,0)



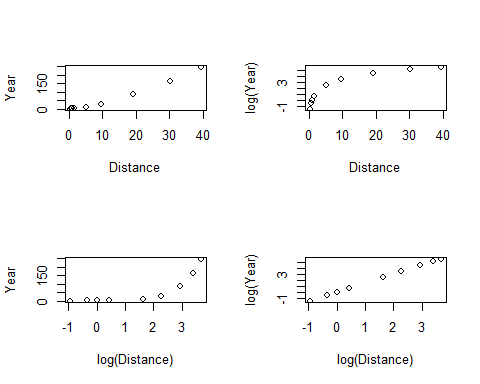
qqnorm(mod2$residuals)  
qqline(mod2$residuals)



plot(mod2, 1:2)

 **What to do when regression assumptions are violated?** *Examples:* 1.Nonlinear patterns in residuals 2.Heteroscedasticity (nonconstant variance) 3.Lack of normality in residuals 4.Outliers: influential points, large residuals

par(mfrow=c(2,2))  
  
plot(Year~Distance, data=Planets)  
plot(log(Year)~Distance, data=Planets)  
plot(Year~log(Distance), data=Planets)  
plot(log(Year)~log(Distance), data=Planets)

 **Data Transformations** *Can be used to:* 1.Address non-linear patterns 2.Stabilize variance 3.Remove skewness from residuals 4.Minimize effects of outliers

**Common Transformations** For either the response (Y) or predictor (X)… - Logarithm: 𝑌→l𝑜𝑔⁡(𝑌) - Square root: 𝑌→√𝑌 - Exponentiation: 𝑌→𝑒^Y - Power function: 𝑌→𝑌^3 - Reciprocal: 𝑌→1/𝑌

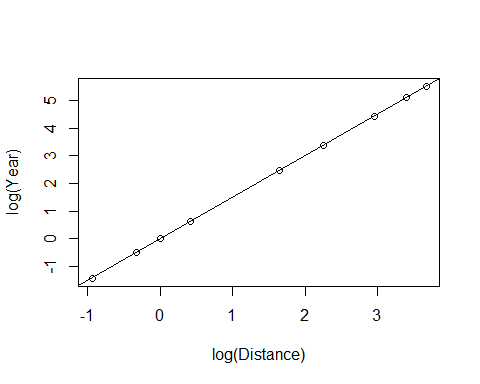
**Example: Planets**

Y = Length of the “year” for planets X = Distance from the sun

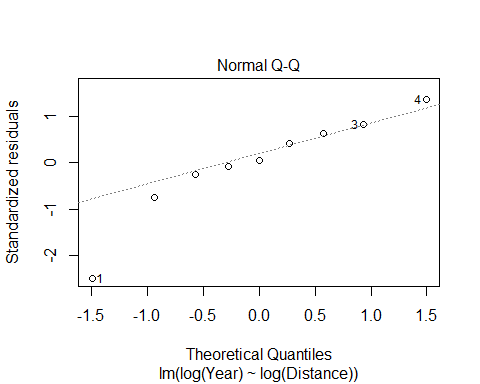
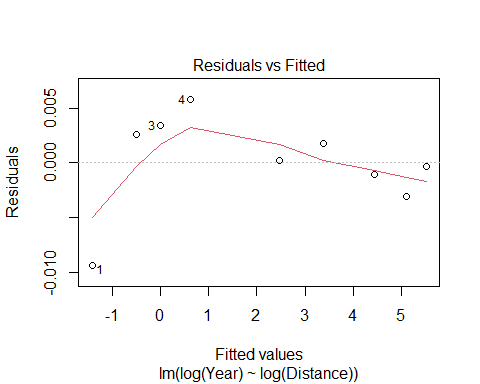
Try scatterplots and LM with Year vs. Distance log(Year) vs. Distance Year vs. log(Distance) log(Year) vs. log(Distance)

*Which transformation gives the best linearity?*

mod3 = lm(log(Year)~log(Distance), data=Planets)  
  
plot(log(Year)~log(Distance), data=Planets)  
abline(mod3)



plot(mod3, 1:2)



summary(mod3)

##   
## Call:  
## lm(formula = log(Year) ~ log(Distance), data = Planets)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.0093289 -0.0010233 0.0002193 0.0025708 0.0057772   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.0034339 0.0020852 -1.647 0.144   
## log(Distance) 1.5020611 0.0009567 1570.016 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.004662 on 7 degrees of freedom  
## Multiple R-squared: 1, Adjusted R-squared: 1   
## F-statistic: 2.465e+06 on 1 and 7 DF, p-value: < 2.2e-16

log(Year) = -0.0034399 + 1.5020611\*log(Distance)

Year = e ^(-0.0034399 + 1.5020611(log(Distance))

Year = e ^(-0.0034399) e ^((1.5020611)(log(Distance))

Year = e ^(-0.0034399) e ^(log(Distance ^1.5020611))

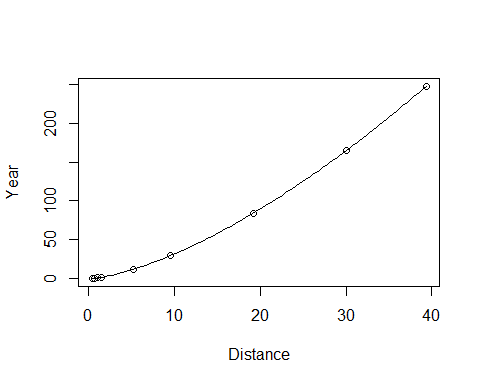
Year = e ^(-0.0034399) (Distance^1.5020611)

exp(-0.0034339)

## [1] 0.996572

Year = 0.996572(Distance^1.5020611)

plot(Year~Distance, data=Planets)  
curve(0.996572\*(x^1.5020611), add=TRUE)



B0 = summary(mod3)$coefficients[1,1]  
B1 = summary(mod3)$coefficients[2,1]  
  
plot(Year~Distance, data=Planets)  
curve(exp(B0)\*x^B1, add=TRUE)

