simple\_reg\_2.R

SANGHOOJEFFREY

Tue Jun 26 19:43:52 2018

# 회귀분석 기초와 상관분석  
# 모형진단  
# 회귀모형은 1)정규성 2)독립성 3)등분산성을 가정하고 있다.  
# 세운 회귀모형이 정규성 독립성 등분산성을 만족하는지 확인  
  
if(!require(car)) install.packages("car"); library(car)

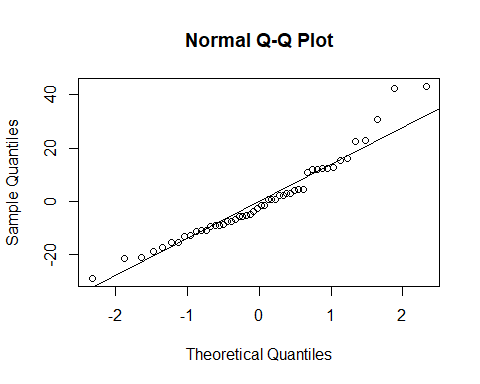
## Loading required package: car

## Warning: package 'car' was built under R version 3.4.4

## Loading required package: carData

## Warning: package 'carData' was built under R version 3.4.4

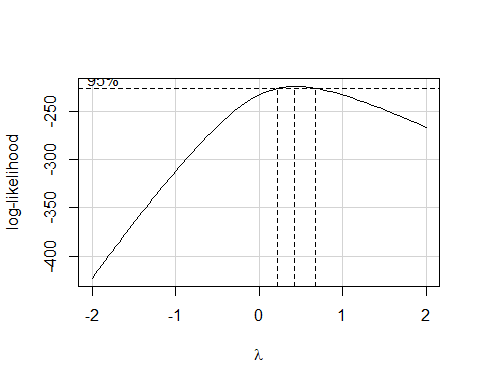
out <- lm(dist~speed, data=cars)  
  
# 1. 정규성  
qqnorm(out$residuals)  
qqline(out$residuals) # Q-Q plot을 이용한 정규성 검정 (점이 선 위에 있을 수록 정규분포를 따름)



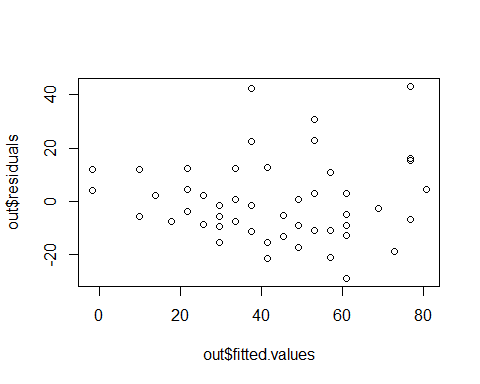
shapiro.test(out$residuals) # 정규성을 만족하지 않음

##   
## Shapiro-Wilk normality test  
##   
## data: out$residuals  
## W = 0.94509, p-value = 0.02152

boxCox(cars$dist~cars$speed) # box-cox 변환을 통해 정규성



# 2. 독립성  
  
plot(out$fitted.values, out$residuals) #잔차도표를 그려서 시각적으로 확인

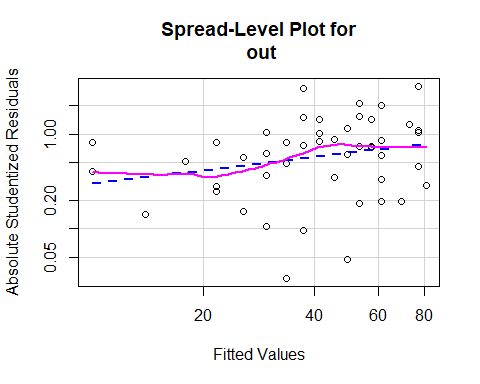


durbinWatsonTest(out)

## lag Autocorrelation D-W Statistic p-value  
## 1 0.1604322 1.676225 0.208  
## Alternative hypothesis: rho != 0

# 3. 등분산성  
spreadLevelPlot(out)

## Warning in spreadLevelPlot.lm(out):   
## 2 negative fitted values removed

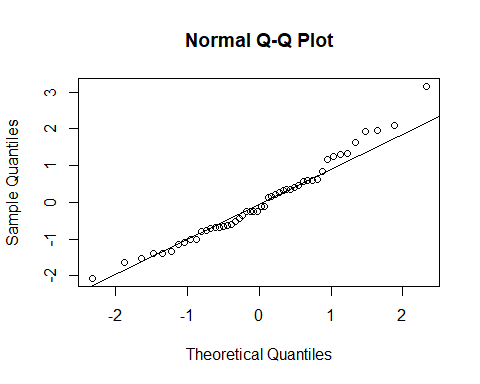


##   
## Suggested power transformation: 0.5402974

ncvTest(out) # 등분산성을 만족하지 않음

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 4.650233 Df = 1 p = 0.03104933

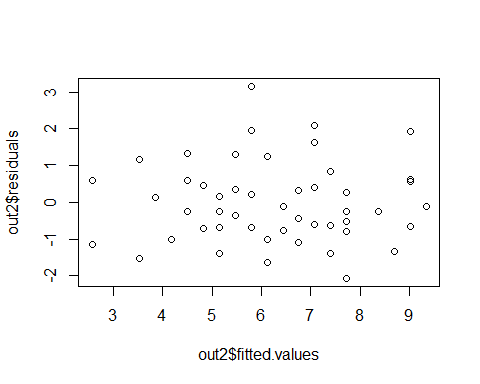
# 정규성과 등분산성을 만족하기 위해 반응변수를 변환  
out2 <- lm(sqrt(dist)~speed,data=cars)  
qqnorm(out2$residuals)  
qqline(out2$residuals) # Q-Q plot을 이용한 정규성 검정 (점이 선 위에 있을 수록 정규분포를 따름)



shapiro.test(out2$residuals) # 정규성을 만족

##   
## Shapiro-Wilk normality test  
##   
## data: out2$residuals  
## W = 0.97332, p-value = 0.3143

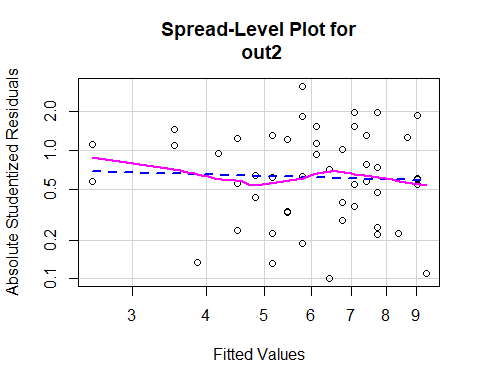
plot(out2$fitted.values, out2$residuals) #잔차도표를 그려서 시각적으로 확인



durbinWatsonTest(out2)

## lag Autocorrelation D-W Statistic p-value  
## 1 0.01762836 1.941736 0.732  
## Alternative hypothesis: rho != 0

spreadLevelPlot(out2)

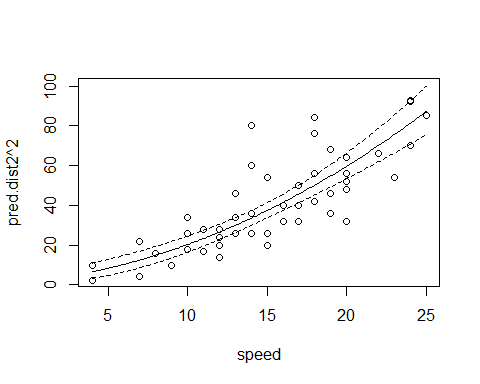


##   
## Suggested power transformation: 1.118288

ncvTest(out2) # 등분산성을 만족

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 0.01205185 Df = 1 p = 0.9125831

# 단순회귀모형의 시각화  
speed <- seq(min(cars$speed), max(cars$speed),.1)  
pred.dist2 <- predict(out2, newdata=data.frame(speed=speed), interval="confidence")  
matplot(speed, pred.dist2^2, type='n')  
matlines(speed, pred.dist2^2, lty=c(1,2,2), col=1) # 선형 회귀식은 직선, 신뢰구간은 점선으로 표현  
matpoints(cars$speed, cars$dist, pch=1)



alligator = data.frame(  
 lnLength = c(3.87, 3.61, 4.33, 3.43, 3.81, 3.83, 3.46, 3.76,  
 3.50, 3.58, 4.19, 3.78, 3.71, 3.73, 3.78),  
 lnWeight = c(4.87, 3.93, 6.46, 3.33, 4.38, 4.70, 3.50, 4.50,  
 3.58, 3.64, 5.90, 4.43, 4.38, 4.42, 4.25)  
)  
  
# Q.5 악어자료의 정규성, 독립성, 등분산성을 확인하세요.  
# 가정을 만족하지 않을 경우 반응변수를 변환시켜보세요  
  
  
# 위의 방법외에도 회귀모형의 결과를 plot()으로 그리면 독립성, 정규성, 등분산성을 시각적으로 확인 가능  
par(mfrow=c(2,2))  
plot(out2)

