Multiple Regression Analysis

yoda

2017년 9월 3일

### 1. 다중회귀분석

단순 회귀 분석의 확장으로 독립 변수(설명 변수)가 2개 이상인 경우를 분석 대상으로 하는 회귀 분석 방법 중 하나이다.

회귀 분석에는 원인이 되는 독립 변수와 결과가 되는 종속 변수가 존재하는데,

이때 종속 변수는 하나이고 독립 변수가 2개 이상인 회귀 모델에 대한 분석을 수행하는 방법이 다중 회귀 분석이다.

본 분석에서는 종속변수를 중고판매가(POSTPRICE)로 설정하고 나머지 컬럼(변수)들을 종속변수를 설명하는

독립변수로 설정을 한다.

분석 후 유의한 회귀식을 도출하여 중고판매가 가격을 예측하고 각 독립변수별 영향력을 살펴본다.

### 2. 데이터 로드 및 확인

data1 <- read.csv("perfact2.csv")  
str(data1)

## 'data.frame': 38912 obs. of 15 variables:  
## $ ID : int 16186605 16186608 16186617 16186621 16186632 16186644 16186673 16186674 16186678 16186679 ...  
## $ POSTDATE : Factor w/ 276 levels "2016-11-30","2016-12-01",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ MODEL : Factor w/ 13 levels "A1303","A1332",..: 10 13 12 8 11 9 4 10 8 8 ...  
## $ GB : Factor w/ 6 levels "128GB","16GB",..: 2 1 4 2 5 5 4 5 5 5 ...  
## $ KRNAME : Factor w/ 13 levels "아이폰 3GS","아이폰 4",..: 9 12 11 7 13 10 4 9 7 7 ...  
## $ POSTPRICE : int 390000 1030000 780000 230000 370000 510000 75000 420000 270000 300000 ...  
## $ CONTRACT : Factor w/ 3 levels "모름","요금할인약정 가능",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ GUARANTEE : Factor w/ 3 levels "무상 보증기간 남음 ",..: 1 1 1 3 1 2 2 3 2 2 ...  
## $ CHANGES : Factor w/ 2 levels "유심기변","확정기변": 2 1 1 1 2 2 2 2 2 2 ...  
## $ CONDITIONS: Factor w/ 4 levels "미사용","상(새제품 수준/무흠집)",..: 2 1 1 2 2 2 3 3 2 3 ...  
## $ COMPONENT : Factor w/ 3 levels "구성품 일부 누락",..: 1 1 1 1 1 1 1 1 3 1 ...  
## $ SOLD : Factor w/ 2 levels "no","yes": 2 1 1 2 1 1 1 2 2 1 ...  
## $ RELDATE : Factor w/ 10 levels "2009-11-28","2010-09-10",..: 7 9 9 6 8 7 4 7 6 6 ...  
## $ RELPRICE : int 869000 1152800 869000 789800 699600 1130800 946000 999900 924000 924000 ...  
## $ FALLRATE : num 55.1 10.7 10.2 70.9 47.1 54.9 92.1 58 70.8 67.5 ...

데이터 개요에서 설명한 데이터를 그대로 사용한다.(총 38912개 관측치, 15개 변수)

### 3. 연속형 변수 타입 변환(as.numeric) 및 표준화

# 연속형 변수의 타입 변환  
data1$POSTDATE <- as.numeric(as.Date(data1$POSTDATE))  
data1$RELPRICE <- as.numeric(data1$RELPRICE)  
data1$FALLRATE <- as.numeric(data1$FALLRATE)  
data1$RELDATE <- as.numeric(as.Date(data1$RELDATE))  
# min, max를 이용하는 표준화 변환 함수 정의  
normal <- function(x){  
 return ((x-min(x)) / (max(x)-min(x)))  
}  
  
# 연속형 변수의 표준화  
data1$RELPRICE <- normal(data1$RELPRICE)  
data1$FALLRATE <- normal(data1$FALLRATE)  
data1$POSTDATE <- normal(data1$POSTDATE)  
data1$RELDATE <- normal(data1$RELDATE)

종속변수 중고판매가격(POSTPRICE)를 설명하는 변수인 POSTDATE, FALLRATE, RELPRICE, RELDATE의 단위가 다르기 때문에 표준화를 해야한다.

표준화방법에는 scale()을 통한 z변환, min값과 max값을 이용하는 [0~1]변환이 있으며 여기서는 [0~1]변환을 사용한다.

먼저 숫자형인 int타입을 num타입으로 변경 후 표준화를 실시한다.

### 4. 연속형 독립변수 간의 상관계수 확인

# 연속형 변수만 따로 분리하여 데이터프레임 생성  
co <- data.frame(data1$RELPRICE, data1$FALLRATE, data1$POSTDATE, data1$RELDATE)  
  
# 변수간 상관계수 확인. 상관계수는 -1 ~ 1 사이의 값을 가지며  
# 0.8 이상이면 다중공선성이 있을 수 있다.  
cor(co)

## data1.RELPRICE data1.FALLRATE data1.POSTDATE data1.RELDATE  
## data1.RELPRICE 1.0000000 -0.288198577 -0.039468703 0.2217497  
## data1.FALLRATE -0.2881986 1.000000000 -0.005323015 -0.8871683  
## data1.POSTDATE -0.0394687 -0.005323015 1.000000000 0.1361389  
## data1.RELDATE 0.2217497 -0.887168325 0.136138944 1.0000000

상관계수 확인 결과 POSTRELGAP~POSTNEXTGAP 과 POSTRELGAP~FALLRATE의 상관계수가 0.8이 넘는 것을 확인할 수 있다.

### 5. 다중회귀분석

names(data1)

## [1] "ID" "POSTDATE" "MODEL" "GB" "KRNAME"   
## [6] "POSTPRICE" "CONTRACT" "GUARANTEE" "CHANGES" "CONDITIONS"  
## [11] "COMPONENT" "SOLD" "RELDATE" "RELPRICE" "FALLRATE"

lm.fit <- lm(POSTPRICE~KRNAME+POSTDATE+GB+CONTRACT+GUARANTEE+CHANGES+CONDITIONS+COMPONENT+SOLD+RELDATE+RELPRICE+FALLRATE  
 , data=data1)  
summary(lm.fit)

##   
## Call:  
## lm(formula = POSTPRICE ~ KRNAME + POSTDATE + GB + CONTRACT +   
## GUARANTEE + CHANGES + CONDITIONS + COMPONENT + SOLD + RELDATE +   
## RELPRICE + FALLRATE, data = data1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -99659 -5111 -41 5016 177072   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 952784.8 5990.6 159.046 < 2e-16  
## KRNAME아이폰 4 133611.6 2968.4 45.011 < 2e-16  
## KRNAME아이폰 4S 317439.6 4369.3 72.653 < 2e-16  
## KRNAME아이폰 5 475355.0 5967.6 79.656 < 2e-16  
## KRNAME아이폰 5C 621257.3 6708.5 92.607 < 2e-16  
## KRNAME아이폰 5S 615810.4 7270.3 84.703 < 2e-16  
## KRNAME아이폰 6 754558.8 8743.0 86.304 < 2e-16  
## KRNAME아이폰 6 플러스 796443.4 10269.4 77.555 < 2e-16  
## KRNAME아이폰 6S 930525.9 10954.2 84.947 < 2e-16  
## KRNAME아이폰 6S 플러스 970756.9 11933.4 81.348 < 2e-16  
## KRNAME아이폰 7 1051801.8 11653.2 90.259 < 2e-16  
## KRNAME아이폰 7 플러스 1149912.1 12793.9 89.880 < 2e-16  
## KRNAME아이폰 SE 911257.8 9678.7 94.151 < 2e-16  
## POSTDATE -6865.9 337.3 -20.357 < 2e-16  
## GB16GB -85525.7 2090.1 -40.920 < 2e-16  
## GB256GB 80958.1 1185.9 68.265 < 2e-16  
## GB32GB -76322.1 1071.6 -71.220 < 2e-16  
## GB64GB -55469.3 1065.8 -52.045 < 2e-16  
## GB8GB 63409.0 4192.6 15.124 < 2e-16  
## CONTRACT요금할인약정 가능 125.6 194.4 0.646 0.51822  
## CONTRACT요금할인약정 불가 681.3 259.9 2.622 0.00875  
## GUARANTEE무상 보증기간 만료 2984.7 264.6 11.280 < 2e-16  
## GUARANTEE미확인 1820.2 323.8 5.620 1.92e-08  
## CHANGES확정기변 -1321.0 229.7 -5.751 8.94e-09  
## CONDITIONS상(새제품 수준/무흠집) -5581.0 245.8 -22.705 < 2e-16  
## CONDITIONS중 (생활흠집 외 양호) -4598.9 311.3 -14.775 < 2e-16  
## CONDITIONS하(번인/잔상/파손) -1000.9 615.2 -1.627 0.10376  
## COMPONENT단품 (기기본체만 있음) 1949.5 397.3 4.907 9.28e-07  
## COMPONENT박스 및 모든 구성품 포함 -369.4 176.6 -2.091 0.03651  
## SOLDyes 821.6 156.0 5.266 1.40e-07  
## RELDATE -1168077.9 12943.9 -90.242 < 2e-16  
## RELPRICE 107090.5 7048.5 15.193 < 2e-16  
## FALLRATE -931097.3 1533.2 -607.309 < 2e-16  
##   
## (Intercept) \*\*\*  
## KRNAME아이폰 4 \*\*\*  
## KRNAME아이폰 4S \*\*\*  
## KRNAME아이폰 5 \*\*\*  
## KRNAME아이폰 5C \*\*\*  
## KRNAME아이폰 5S \*\*\*  
## KRNAME아이폰 6 \*\*\*  
## KRNAME아이폰 6 플러스 \*\*\*  
## KRNAME아이폰 6S \*\*\*  
## KRNAME아이폰 6S 플러스 \*\*\*  
## KRNAME아이폰 7 \*\*\*  
## KRNAME아이폰 7 플러스 \*\*\*  
## KRNAME아이폰 SE \*\*\*  
## POSTDATE \*\*\*  
## GB16GB \*\*\*  
## GB256GB \*\*\*  
## GB32GB \*\*\*  
## GB64GB \*\*\*  
## GB8GB \*\*\*  
## CONTRACT요금할인약정 가능   
## CONTRACT요금할인약정 불가 \*\*   
## GUARANTEE무상 보증기간 만료 \*\*\*  
## GUARANTEE미확인 \*\*\*  
## CHANGES확정기변 \*\*\*  
## CONDITIONS상(새제품 수준/무흠집) \*\*\*  
## CONDITIONS중 (생활흠집 외 양호) \*\*\*  
## CONDITIONS하(번인/잔상/파손)   
## COMPONENT단품 (기기본체만 있음) \*\*\*  
## COMPONENT박스 및 모든 구성품 포함 \*   
## SOLDyes \*\*\*  
## RELDATE \*\*\*  
## RELPRICE \*\*\*  
## FALLRATE \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14260 on 38879 degrees of freedom  
## Multiple R-squared: 0.9967, Adjusted R-squared: 0.9967   
## F-statistic: 3.698e+05 on 32 and 38879 DF, p-value: < 2.2e-16

종속변수를 중고판매가(SOLDPRICE)로 설정하고 회귀분석 실시한 결과 p-value: < 2.2e-16로 회귀모델은 유의하다고 판단할 수 있다.

또한 Multiple R-squared가 0.9975로 독립변수들이 종속변수들을 99.75%만큼 설명할 수 있다.

쉽게 판단하기 위해서 해당 독립변수가 유의하면 Pr(>|t|) 뒤에 '*' 표시가 붙으며 '*\*\*'시가 많을수록 통계적으로 유의할 확률이 크다는 뜻이다.

하지만 Multiple R-squared가 1에 가까운 것에 비해 상대적으로 계수들의 검정통계량이 작아 다중공선성을 의심해야 한다.

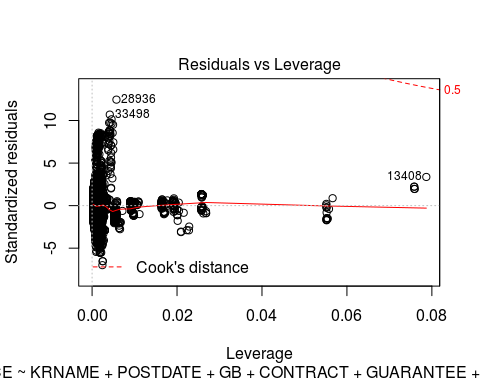
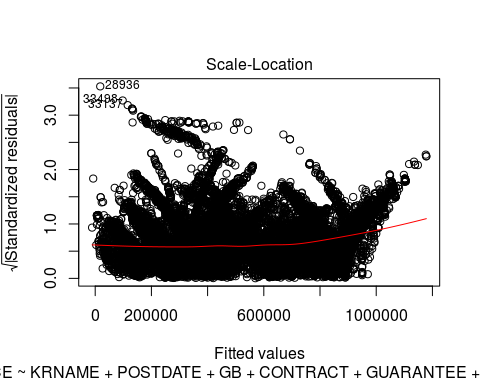
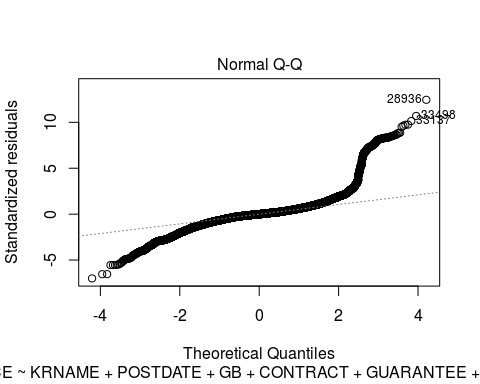
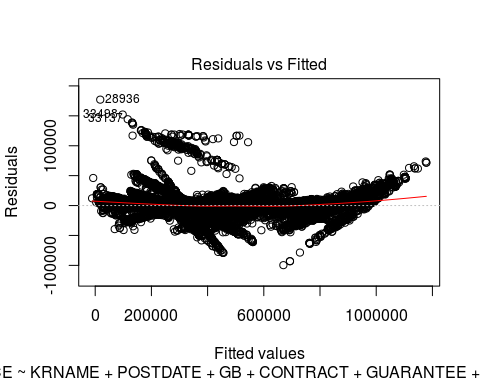
# 회귀분석시 독립변수들의 형식확인  
head(model.matrix(lm.fit), 2)

## (Intercept) KRNAME아이폰 4 KRNAME아이폰 4S KRNAME아이폰 5  
## 1 1 0 0 0  
## 2 1 0 0 0  
## KRNAME아이폰 5C KRNAME아이폰 5S KRNAME아이폰 6 KRNAME아이폰 6 플러스  
## 1 0 0 0 0  
## 2 0 0 0 0  
## KRNAME아이폰 6S KRNAME아이폰 6S 플러스 KRNAME아이폰 7  
## 1 1 0 0  
## 2 0 0 0  
## KRNAME아이폰 7 플러스 KRNAME아이폰 SE POSTDATE GB16GB GB256GB GB32GB  
## 1 0 0 0 1 0 0  
## 2 1 0 0 0 0 0  
## GB64GB GB8GB CONTRACT요금할인약정 가능 CONTRACT요금할인약정 불가  
## 1 0 0 0 0  
## 2 0 0 0 0  
## GUARANTEE무상 보증기간 만료 GUARANTEE미확인 CHANGES확정기변  
## 1 0 0 1  
## 2 0 0 0  
## CONDITIONS상(새제품 수준/무흠집) CONDITIONS중 (생활흠집 외 양호)  
## 1 1 0  
## 2 0 0  
## CONDITIONS하(번인/잔상/파손) COMPONENT단품 (기기본체만 있음)  
## 1 0 0  
## 2 0 0  
## COMPONENT박스 및 모든 구성품 포함 SOLDyes RELDATE RELPRICE FALLRATE  
## 1 0 1 0.8020097 0.5413625 0.5483871  
## 2 0 0 0.9374767 0.8552311 0.1008065

범주형 변수들은 자동적으로 더미변수로 처리하여 분석을 실행하는 것을 볼 수 있다.

### 6. 잔차 확인

plot(lm.fit)

 Residuals VS Fitted plot을 보면 일정한 패턴을 보이며 0을 기준으로 고루 퍼져있지 않다. - 등분산성 위배

Normal Q-Q plot 을 보면 직선형태가 아닌 S자 형태를 보인다. - 정규성 위배

### 7. 다중공선성 확인

# library(car)  
library(pedometrics)  
stepVIF(lm.fit)

##   
## Call:  
## lm(formula = POSTPRICE ~ KRNAME + POSTDATE + GB + CONTRACT +   
## GUARANTEE + CHANGES + CONDITIONS + COMPONENT + SOLD, data = data1)  
##   
## Coefficients:  
## (Intercept) KRNAME아이폰 4   
## 317643 16777   
## KRNAME아이폰 4S KRNAME아이폰 5   
## 31540 42915   
## KRNAME아이폰 5C KRNAME아이폰 5S   
## 86357 106758   
## KRNAME아이폰 6 KRNAME아이폰 6 플러스   
## 191158 267832   
## KRNAME아이폰 6S KRNAME아이폰 6S 플러스   
## 302494 376056   
## KRNAME아이폰 7 KRNAME아이폰 7 플러스   
## 501065 629999   
## KRNAME아이폰 SE POSTDATE   
## 218613 -97254   
## GB16GB GB256GB   
## -131741 90035   
## GB32GB GB64GB   
## -98891 -63474   
## GB8GB CONTRACT요금할인약정 가능   
## -137690 5539   
## CONTRACT요금할인약정 불가 GUARANTEE무상 보증기간 만료   
## 3799 -10697   
## GUARANTEE미확인 CHANGES확정기변   
## -9999 -11735   
## CONDITIONS상(새제품 수준/무흠집) CONDITIONS중 (생활흠집 외 양호)   
## -74387 -101315   
## CONDITIONS하(번인/잔상/파손) COMPONENT단품 (기기본체만 있음)   
## -146959 17368   
## COMPONENT박스 및 모든 구성품 포함 SOLDyes   
## 18346 -6057

stepVIF로 다중공선성을 띄고 있는 변수들을 제거한 결과 유의한 변수는

POSTPRICE ~ KRNAME + POSTDATE + GB + CONTRACT + GUARANTEE + CHANGES + CONDITIONS + COMPONENT + SOLD로 나타났다.

### 8. 다중공선성이 있는 변수 제거 후 다항회귀분석

lm.fit2 <- lm(formula(stepVIF(lm.fit)), data=data1)

### 9. 최적의 회귀분석모델을 얻기 위해 STEPWISE 실시

lm.fit.step <- step(lm.fit2, direction="both")

## Start: AIC=835938.8  
## POSTPRICE ~ KRNAME + POSTDATE + GB + CONTRACT + GUARANTEE + CHANGES +   
## CONDITIONS + COMPONENT + SOLD  
##   
## Df Sum of Sq RSS AIC  
## <none> 8.3038e+13 835939  
## - CONTRACT 2 1.6675e+11 8.3205e+13 836013  
## - SOLD 1 3.0818e+11 8.3346e+13 836081  
## - GUARANTEE 2 3.7594e+11 8.3414e+13 836111  
## - CHANGES 1 5.3595e+11 8.3574e+13 836187  
## - COMPONENT 2 2.4253e+12 8.5463e+13 837055  
## - POSTDATE 1 2.1881e+13 1.0492e+14 845038  
## - CONDITIONS 3 3.4280e+13 1.1732e+14 849380  
## - GB 5 5.6650e+13 1.3969e+14 856167  
## - KRNAME 12 3.5019e+14 4.3322e+14 900196

최종적으로 선택된 설명변수는 다음과 같다.

formula(lm.fit.step)

## POSTPRICE ~ KRNAME + POSTDATE + GB + CONTRACT + GUARANTEE + CHANGES +   
## CONDITIONS + COMPONENT + SOLD

### 10. STEPWISE를 통해 최종적으로 선택된 변수로 최상의 회귀분석 실행

lm.fit.best <- lm(formula(lm.fit.step), data=data1)  
summary(lm.fit.best)

##   
## Call:  
## lm(formula = formula(lm.fit.step), data = data1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -330863 -26864 -40 26536 324895   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 317642.8 7492.7 42.394 < 2e-16  
## KRNAME아이폰 4 16777.1 8512.4 1.971 0.048741  
## KRNAME아이폰 4S 31540.2 9143.1 3.450 0.000562  
## KRNAME아이폰 5 42915.1 7605.7 5.642 1.69e-08  
## KRNAME아이폰 5C 86357.4 9744.5 8.862 < 2e-16  
## KRNAME아이폰 5S 106757.6 7448.2 14.333 < 2e-16  
## KRNAME아이폰 6 191158.3 7396.0 25.846 < 2e-16  
## KRNAME아이폰 6 플러스 267831.7 7443.5 35.982 < 2e-16  
## KRNAME아이폰 6S 302494.4 7413.4 40.803 < 2e-16  
## KRNAME아이폰 6S 플러스 376055.7 7443.4 50.522 < 2e-16  
## KRNAME아이폰 7 501065.0 7433.7 67.404 < 2e-16  
## KRNAME아이폰 7 플러스 629998.5 7449.2 84.573 < 2e-16  
## KRNAME아이폰 SE 218613.2 7480.7 29.224 < 2e-16  
## POSTDATE -97254.0 960.8 -101.222 < 2e-16  
## GB16GB -131741.0 1239.1 -106.316 < 2e-16  
## GB256GB 90034.6 1858.5 48.444 < 2e-16  
## GB32GB -98890.7 857.2 -115.367 < 2e-16  
## GB64GB -63474.2 1241.3 -51.137 < 2e-16  
## GB8GB -137689.6 12045.7 -11.431 < 2e-16  
## CONTRACT요금할인약정 가능 5538.7 629.0 8.806 < 2e-16  
## CONTRACT요금할인약정 불가 3798.5 842.1 4.511 6.48e-06  
## GUARANTEE무상 보증기간 만료 -10697.4 836.7 -12.786 < 2e-16  
## GUARANTEE미확인 -9998.5 1040.3 -9.611 < 2e-16  
## CHANGES확정기변 -11734.7 740.8 -15.842 < 2e-16  
## CONDITIONS상(새제품 수준/무흠집) -74386.7 704.8 -105.539 < 2e-16  
## CONDITIONS중 (생활흠집 외 양호) -101315.4 863.3 -117.352 < 2e-16  
## CONDITIONS하(번인/잔상/파손) -146958.9 1834.0 -80.131 < 2e-16  
## COMPONENT단품 (기기본체만 있음) 17368.0 1283.7 13.530 < 2e-16  
## COMPONENT박스 및 모든 구성품 포함 18345.6 563.1 32.581 < 2e-16  
## SOLDyes -6057.3 504.2 -12.013 < 2e-16  
##   
## (Intercept) \*\*\*  
## KRNAME아이폰 4 \*   
## KRNAME아이폰 4S \*\*\*  
## KRNAME아이폰 5 \*\*\*  
## KRNAME아이폰 5C \*\*\*  
## KRNAME아이폰 5S \*\*\*  
## KRNAME아이폰 6 \*\*\*  
## KRNAME아이폰 6 플러스 \*\*\*  
## KRNAME아이폰 6S \*\*\*  
## KRNAME아이폰 6S 플러스 \*\*\*  
## KRNAME아이폰 7 \*\*\*  
## KRNAME아이폰 7 플러스 \*\*\*  
## KRNAME아이폰 SE \*\*\*  
## POSTDATE \*\*\*  
## GB16GB \*\*\*  
## GB256GB \*\*\*  
## GB32GB \*\*\*  
## GB64GB \*\*\*  
## GB8GB \*\*\*  
## CONTRACT요금할인약정 가능 \*\*\*  
## CONTRACT요금할인약정 불가 \*\*\*  
## GUARANTEE무상 보증기간 만료 \*\*\*  
## GUARANTEE미확인 \*\*\*  
## CHANGES확정기변 \*\*\*  
## CONDITIONS상(새제품 수준/무흠집) \*\*\*  
## CONDITIONS중 (생활흠집 외 양호) \*\*\*  
## CONDITIONS하(번인/잔상/파손) \*\*\*  
## COMPONENT단품 (기기본체만 있음) \*\*\*  
## COMPONENT박스 및 모든 구성품 포함 \*\*\*  
## SOLDyes \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 46210 on 38882 degrees of freedom  
## Multiple R-squared: 0.9656, Adjusted R-squared: 0.9656   
## F-statistic: 3.763e+04 on 29 and 38882 DF, p-value: < 2.2e-16

최종적으로 회귀분석 실시한 결과 p-value: < 2.2e-16로 회귀모델은 유의하다고 판단할 수 있다.

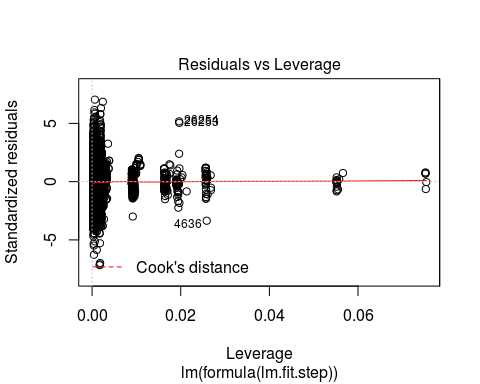
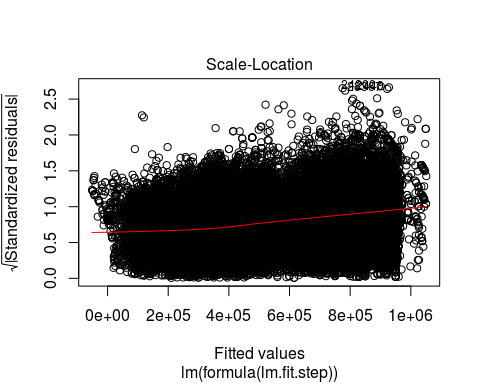
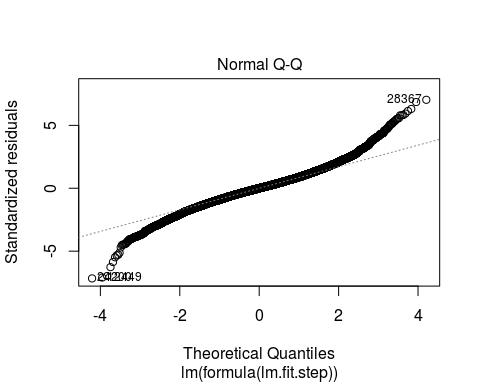
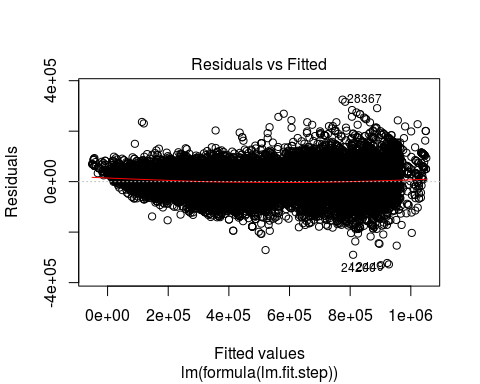
또한 Multiple R-squared가 0.9975로 독립변수들이 종속변수들을 99.75%만큼 설명할 수 있다.

도출된 회귀식은

Y(중고판매가격) = 317642.8 + 16777.1(아이폰 4) + 31540.2(아이폰 4S) + 42915.1(아이폰 5) + 86357.4(아이폰 5C) + 106757.6(아이폰 5S) + 191158.3(아이폰 6) + 267831.7(아이폰 6 플러스) + 302494.4(아이폰 6S) + 376055.7(아이폰 6S 플러스) + 501065(아이폰 7) + 629998.5(아이폰 7 플러스) + 218613.2(아이폰 SE) - 97254(POSTDATE) - 131741(16GB) + 90034.6(256GB) - 98890.7(32GB) - 63474.2(64GB) - 137689.6(8GB) + 5538.7(요금할인약정 가능) + 3798.5(요금할인약정 불가) - 10697.4(무상 보증기간 만료) -9998.5(미확인) - 11734.7(확정기변) - 74386.7(CONDITION 상) - 101315.4(CONDITION 중) - 146958.9(CONDITION 하) + 17368.0(기기본체만 있음) + 18345.6(모든 구성품 포함) - 6057.3(SOLD YES)

### 11. 잔차 확인

plot(lm.fit.best)

 초기 회귀분석과는 다르게 이상적인 모습은 아니지만 잔차가 정규성과 등분산성을 어느정도 띄고 있는 것을 확인할 수 있다.

### 12. 잔차 검증

# 독립성 검증  
library(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

dwtest(lm.fit.best)

##   
## Durbin-Watson test  
##   
## data: lm.fit.best  
## DW = 1.849, p-value < 2.2e-16  
## alternative hypothesis: true autocorrelation is greater than 0

# 정규성 검증  
shapiro.test(sample(resid(lm.fit.best), 5000))

##   
## Shapiro-Wilk normality test  
##   
## data: sample(resid(lm.fit.best), 5000)  
## W = 0.98618, p-value < 2.2e-16

독립성 검증인 Durbin-Watson test 경우 p-value < 2.2e-16로 자기상관을 띈다고도 볼 수 있지만 해석에 따라 DW값이 2에 가까우면 독립성을 만족하다고 볼 수 도 있다.

정규성 검증인 Shapiro-Wilk normality test 경우 p-value < 2.2e-16로 잔차가 정규성을 만족하지 않는 것을 확인 할 수 있다.

이는 표본 수가 부족하거나 데이터에 이상치들이 존재한다고 생각 할 수 있다.

# tail(model.matrix(lm1))

# lm1.step <- step(lm1, direction="both")  
# formula(lm1.step)

# multi.best.lm1 <- lm(formula(lm1.step), data=df)  
# summary(multi.best.lm1)