다중회귀분석을 이용한 항공기 가격 예측

daja Vu 통계학회 3 조 프로젝트 발표 소논문

2020.11.12

20190779 정다인

20190781 정솔잎

20180768 박주연

20180769 박혜연

20180777 이가은

목차

제 1 장 서론

1.1 주제선정 동기

요즘 테슬라, 자율주행자동차와 같은 자동차에 관한 이야기가 이슈다. 주제를 알아보는 중 기사를 통해 자동차 가격에 영향을 주는 요인을 알게 되었고, 조금은 생소한 비행기가격은 어떻게 결정될지 궁금증이 생겨 주제로 결정하게 되었다. 일방적으로 우리가 여행갈때 타는 비행기를 포함하여, 전투기, 민간수송비행기 등 앞으로 사용할 데이터는 외국기업에서 제조되는 항공기에 관한 데이터를 다룰 것이다.

지금부터는 다중회귀분석을 통해 항공기가격에 영향을 주는 요인을 소개하고자 한다.

```
#경로지정
setwd("C:/Users/user/Desktop/스터디")
```

제 2 장 데이터 수집 및 처리

2.1 데이터 수집

앞으로 사용할 데이터는 공공데이터포털 (https://www.data.go.kr)에서 제공하는 국토교통부_세계항공기 정보.csv 파일이다. 총 296 행 18 열 형태로 열이름은 제조사, 비행기모델, IACO CODE, IATA CODE, 비행기구분, 비행기크기, class, 엔진타입, 엔진수, 최초운항일, 생산수량, 단가.백만달라, 상태, 조종석수, 승객수, 길기.m, 높이.m,운송범위.km 로 이뤄져있다.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	제조사	비행기모델	ICAO COD	IATA COD	비행기구분	비행기크기	CLASS	엔진타입	엔진수	최초운항일	생산수량	단가	상태	조종석수	승객수	길이(m	높이(m	운송범위(k
2	CURTISS	Curtiss C-46	C46	CWC	LandPlane	MEDIUM	Military tra	Piston	2	1940-03-26	3181		Produc	5	40	23.3	6.6	870
3	CONVAIR	Convair CV-	CVLP	CV4	LandPlane	MEDIUM	Airliner	Piston	2	1947-03-16	1181		Produc	3	40	22.8	8.2	1900
4	LOCKHEE	Lockheed L-	CONI	L49	LandPlane	MEDIUM	Airliner	Piston	4	1951-07-14	259		Produc	5	106	34.6	7.5	8288
5	DOUGLA	Douglas DC	DC6	D6F	LandPlane	MEDIUM	Airliner/tra	Piston	4	1946-02-15	704		Produc	4	68	30.7	8.7	7377
6	DE HAVI	DHC-2 Turb	DH2T	DHR	LandPlane	LOW	STOLutility	Turboprop	1	1947-08-16	1657		Produc	1	6	9.2	2.7	732
7	DE HAVI	DHC-2 Beav	DHC2	DHP	LandPlane	LOW	STOLutility	Piston	1	1947-08-16	1657		Produc	1	6	9.2	2.7	732
8	DE HAVI	DHC-3 Otte	DHC3	DHL	LandPlane	LOW	STOL utilit	Piston	1	1951-12-12	466	0.1	Produc	2	11	12.8	3.8	1524
9	DE HAVI	DH.104 Dov	DOVE	DHD	LandPlane	LOW	short-haul	Piston	2	1945-09-25	544	0.1	Produc	2	8	12	4.1	1420
10	DE HAVI	DHC-4 Carit	DHC4	DHC	LandPlane	MEDIUM	STOL Tran	Piston	2	1958-07-30	307		Produc	2	30	22.1	9.7	2104
11	GULFSTR	Aerospace G	G159	GRS	LandPlane	MEDIUM	Business a	Turboprop	2	1958-08-14	200		Produc	2	24	19.4	6.9	4090
12	SIKORSK	Sikorsky S-	S58T	S58	Helicopter	LOW	Helicopter	Turboprop	1	1954-03-08	2108		Produc	2	12	17.3	4.9	293
13	DOUGLA	Douglas DC	DC85	D8T	LandPlane	HIGH	Narrow-bc	Jet	4	1958-05-30			Produc	2	189	45.9		10843
14	DOUGLA	Douglas DC	DC86	D8L	LandPlane	HIGH	Narrow-bc	Jet	4				Produc	2	189	48		9600
15	DOUGLA	Douglas DC	DC87	D8Q	LandPlane	HIGH	Narrow-bc	Jet	4				Produc	2	189	48		9600

2.2 데이터 처리

데이터를 불러와 확인해본 결과 결측치가 존재하였고 데이터를 보다 쉽게 처리하기 위해 변수이름을 변경한다. 또한 범주형 자료도 factor 처리한다.

```
# 열 이름 변경
colnames(worldflights) = c("제조사","비행기모델","ICAO","IATA",
"비행기구분","비행기크기","class","엔진타입","엔진수","최초운항일",
                           "생산수량","price","상태","조종석수","승객수","길이
","높이","운송범위")
colnames(worldflights)
```

```
##  [1] "제조사"     "비행기모델" "ICAO"       "IATA"       "비행기구분"
##  [6] "비행기크기" "class"      "엔진타입"   "엔진수"     "최초운항일"
## [11] "생산수량"   "price"      "상태"       "조종석수"   "승객수"
## [16] "길이"       "높이"       "운송범위"
```

종속변수(price)와 총 17 개의 독립변수 중 의미 있다고 생각하는 독립변수 12 개 선택

```
library(dplyr)wf2 <- select(worldflights, "제조사", "비행기구분","비행기크기
","class","엔진타입",
              "엔진수", "생산수량","price","조종석수","승객수",
              "길이","높이","운송범위")
```

종속변수인 price 의 데이터타입이 문자형으로 출력되어 숫자형으로 변경해준다.

```
as.data.frame(wf2)

is.numeric(wf2$price)

## [1] FALSE

wf2$price=as.numeric(wf2$price)

is.numeric(wf2$price)

## [1] TRUE
```

결측치가 포함되면 회귀분석 진행이 어려우므로 가격, 승객수, 길이, 높이, 운송범위, 생산수량 에서 결측치가 포함된 행은 모두 제거한다.

```
wf2=wf2%>%
  filter(price != 0)

wf3=wf2%>%
  filter(승객수 != 0)

wf4=wf3%>%
  filter(길이 != 0)

wf5=wf4%>%
  filter(높이 != 0)

wf6=wf5%>%
  filter(운송범위 != 0)

wf7=wf6%>%
  filter(생산수량 != 0)

dim(wf7)

## [1] 136  13

as.data.frame(wf7)
```

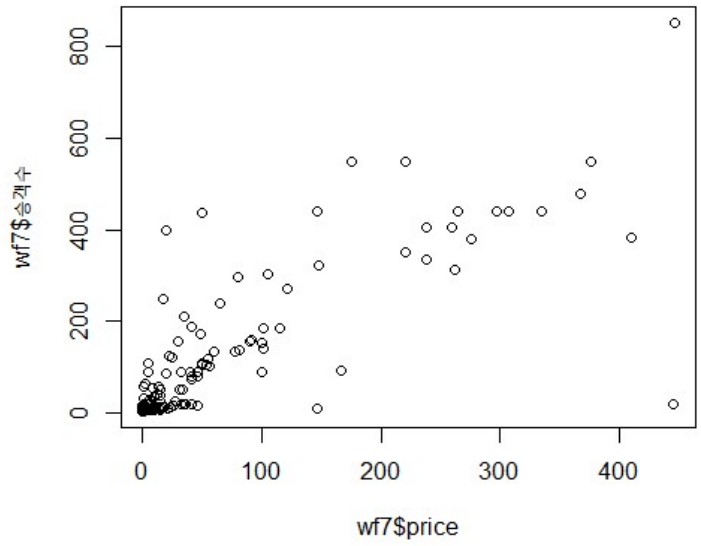
제조사, 비행기크기, 비행기 구분, 엔진타입은 범주형 변수이기 때문에 factor 로 처리한다.

```
wf7$제조사 = factor(wf7$제조사)


wf7$비행기크기 = factor(wf7$비행기크기)


wf7$비행기구분= factor(wf7$비행기구분)


wf7$엔진타입 = factor(wf7$엔진타입
```

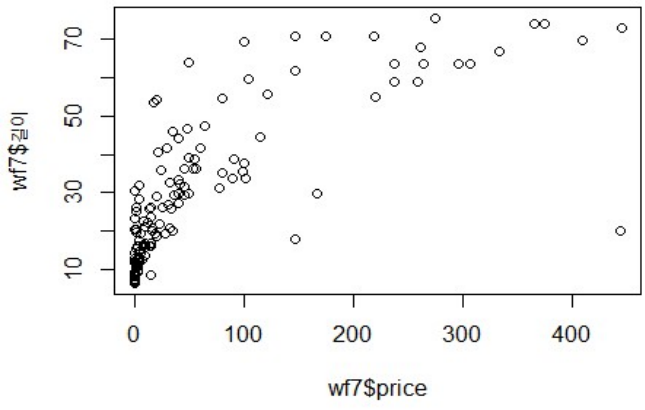
2.3 데이터 특성파악

가격과 상관계수가 높은 4 가지 변수 승객수, 길이, 높이, 운송범위 와의 산점도

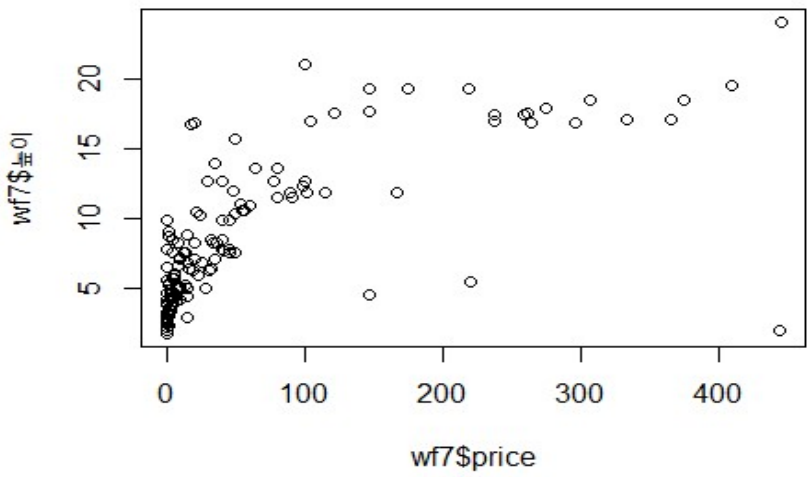
```
par(mar=c(4.5,4.5,0.5,0.5))
plot(wf7$price, wf7$승객수)
```



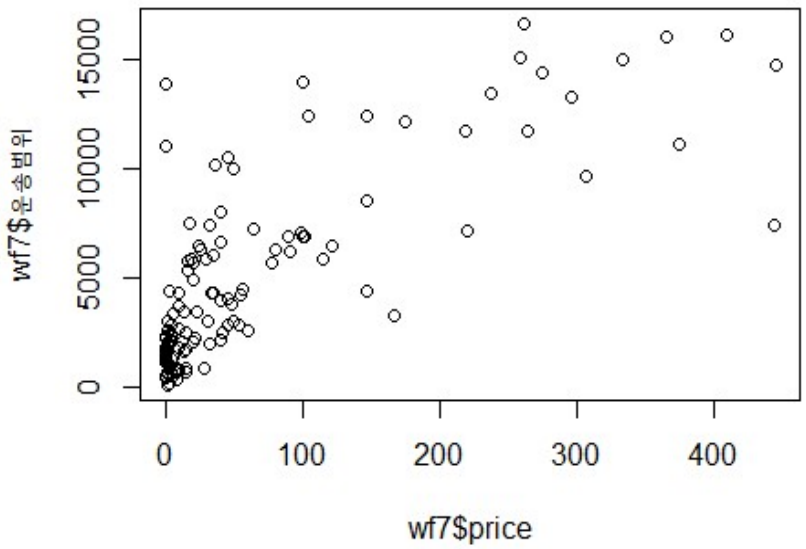

```
plot(wf7$price, wf7$길이)
```



```
plot(wf7$price, wf7$높이)
```




```
plot(wf7$price, wf7$운송범위)
```



제 3 장 실증 분석

여러개의 독립변수를 가지므로 다중회귀분석을 실행한다. 종속변수는 가격이고, 독립변수는 제조사, 비행기구분, 비행기 크기, class, 엔진타입, 엔진 수, 생산수량, 조종석 수, 승객 수, 길이, 높이, 운송범위이다.

3.1 다중회귀모형 적합

회귀분석을 수행하기 위한 가설을 설정한다. 귀무가설: 모든 독립변수는 price 에 영향을 미치지 않을 것이다. 대립가설: price 에 적어도 하나는 영향을 미칠 것이다.

```
fit <- lm(price~., data = wf7)
summary(fit)

## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56.58 on 37 degrees of freedom
## Multiple R-squared:  0.9144,  Adjusted R-squared:  0.6876
## F-statistic: 4.032 on 98 and 37 DF,  p-value: 4.068e-06
```

회귀모형을 적합시키고 F-test 결과를 확인하면 P-value 가 매우 작기 때문에 귀무가설을 기각한다. 즉, 적어도 하나의 독립변수는 종속변수에 유의한 영향을 미친다고 볼 수 있다.

3.1.1 변수 선택

회귀모형은 유의하게 나왔으나 유의하지 않은 변수가 많아 최적의 회귀방정식을 선택하기 위해 단계적 변수선택 방법을 이용한다. 3 개의 방법(전진선택법, 후진제거법, 단계선택법)을 각각 모형에 적용해보고 분석 데이터에 가장 잘 맞는 모형을 찾아내고자 한다.

1. 전진선택법

```
    fit.con1 <- lm(price~1,data = wf7)
    fit.forward1 <- step(fit.con1,scope=list(lower=fit.con1,upper=fit),direct
    ion = "forward")

    ## ---
    ## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
    ##
    ## Residual standard error: 52.4 on 132 degrees of freedom
```

```
## Multiple R-squared:  0.738,    Adjusted R-squared:  0.7321
## F-statistic:   124 on 3 and 132 DF,  p-value: < 2.2e-16
```

선택된 최종 변수는 승객수, 운송범위, 높이이며 수정된 결정계수는 0.7321 임을 확인할 수 었다.

2. 후진제거법

```
fit.backward <- step(fit, scope = list(lower = fit.con1, upper = fit),dir
ection = "backward")

summary(fit.backward)

##
## Call:
## lm(formula = price ~ 제조사 + class + 엔진수 + 생산수량 + 길이 +
##     높이 + 운송범위, data = wf7)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -120.1    0.0    0.0    0.0  225.7
##
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.17 on 41 degrees of freedom
## Multiple R-squared:  0.9131,  Adjusted R-squared:  0.7137
## F-statistic: 4.581 on 94 and 41 DF,  p-value: 2.496e-07
```

선택된 최종 변수는 제조사, class, 엔진수, 생산수량, 길이, 높이, 운송범위이며 수정된 결정계수는 0.7137 임을 확인할 수 있다.

3. 단계선택법

```
fit.both <- step(fit.con1, scope = list(lower = fit.con1, upper = fit), d
irection = "both")

summary(fit.both)

##
## Call:
## lm(formula = price ~ 승객수 + 운송범위 + 높이, data = wf7)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -176.88  -15.41   -0.85   16.25  348.14
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.692479  11.138049   1.140 0.256533
## 승객수       0.470124   0.061213   7.680 3.17e-12 ***
## 운송범위     0.011942   0.001781   6.707 5.30e-10 ***
## 높이        -7.019925   2.028752  -3.460 0.000727 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 52.4 on 132 degrees of freedom
## Multiple R-squared:  0.738,   Adjusted R-squared:  0.7321
## F-statistic:   124 on 3 and 132 DF,  p-value: < 2.2e-16
```

선택된 최종 변수는 승객수, 운송범위, 높이이며 수정된 결정계수는 0.7321 임을 확인할 수 있었다.

전진선택법과 단계선택법의 최종 결과가 같게 나와 단계선택법을 적용한 모형을

최종 모형으로 결정한다.

3.2 회귀모형 진단

3.2.1 회귀계수 진단

t-test 를 이용해 승객수, 운송범위, 높이의 회귀계수의 유의성을 검정한 결과 세 독립변수의 유의확률이 모두 매우 작게 나와 모든 회귀계수가 유의한 영향을 미친다는 것을 알 수 있다.

```
summary(fit.both)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.692479  11.138049   1.140 0.256533
## 승객수       0.470124   0.061213   7.680 3.17e-12 ***
## 운송범위     0.011942   0.001781   6.707 5.30e-10 ***
## 높이        -7.019925   2.028752  -3.460 0.000727 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 52.4 on 132 degrees of freedom
## Multiple R-squared:  0.738,   Adjusted R-squared:  0.7321
## F-statistic:   124 on 3 and 132 DF,  p-value: < 2.2e-16
```

3.2.2 회귀모형의 설명력

회귀모형은 이고 수정된 결정계수는 0.7321 이므로 회귀 직선이 자료의 73.21%를 설명함을 알 수 있다.

3.2.3 다중공선성 확인

독립변수들 간의 상관관계가 있는지 알아보기 위해 다중공선성 함수인 vif()를 이용해 확인하였다. 그 결과, 승객수는 4.665014, 운송범위는 2.918100, 높이는 5.583548 로 모두 10 미만이므로 독립변수들 사이에 심각한 선형관계는 없다고 할 수 있다.

```
library(car)

vif(fit.both)

##    승객수  운송범위      높이
## 4.665014 2.918100 5.583548
```

3.2.4 선형회귀모형의 4 가지 기본가정

1. 독립성

더빈-왓슨 통계량을 이용해 오차항들 간의 자기상관 여부를 검정한다. DW 가 1.6859 로 2 에 가까운 값을 가지므로 자기상관에 문제가 없다고 판단했다.

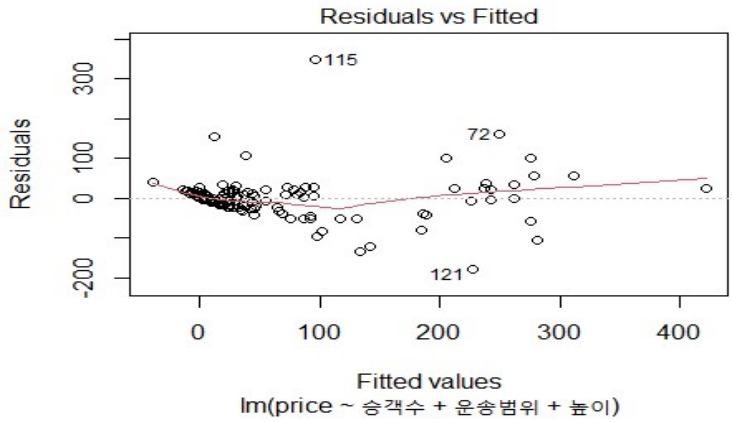
```
#모형의 잔차 독립성을 확인해 주는 더빈왓슨통계량
library(lmtest)

dwtest(fit.both)

##
##   Durbin-Watson test
##
## data:  fit.both
## DW = 1.6859, p-value = 0.02555
## alternative hypothesis: true autocorrelation is greater than 0
```

1.6859 2 에 가까운값을 가지며 독립변수 잔차들 간의 자기상관이 없다

2. 선형성

예측값(fitted)과 잔차(residual)을 비교한 선형성 그래프를 확인해본 결과, 잔차의 추세가 0 에서 크게 벗어나지 않으므로 예측값에 따라 잔차가 크게 달라지지 않는다고 판단하였다.

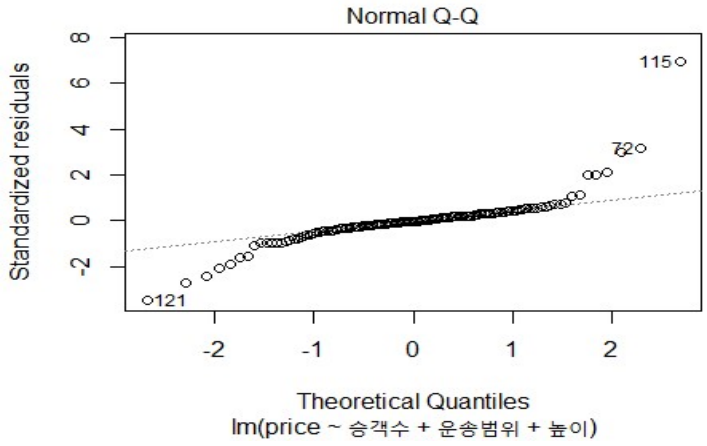
```
plot(fit.both,which=1)
```



3. 정규성

Normal Q-Q 그래프 결과, 정규성을 따른다는 것을 확인할 수 있었다. 또한, 샤피로의 검정으로 확인한 결과 유의확률이 매우 작으므로 잔차의 정규성이 위반되지 않는다고 판단하였다.

```
plot(fit.both,which=2)
```



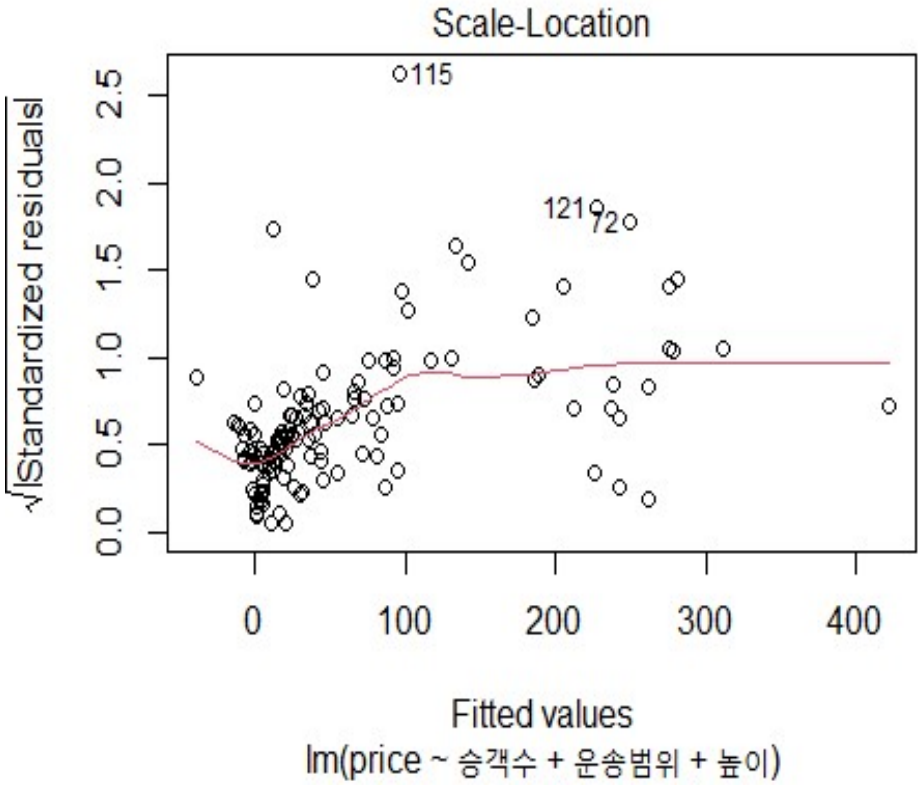
정규성 확인 할 수 있는 샤피로의 검정

```
shapiro.test(fit.both$residuals)

##
##   Shapiro-Wilk normality test
##
## data:  fit.both$residuals
## W = 0.7593, p-value = 1.175e-13
```

4. 등분산성

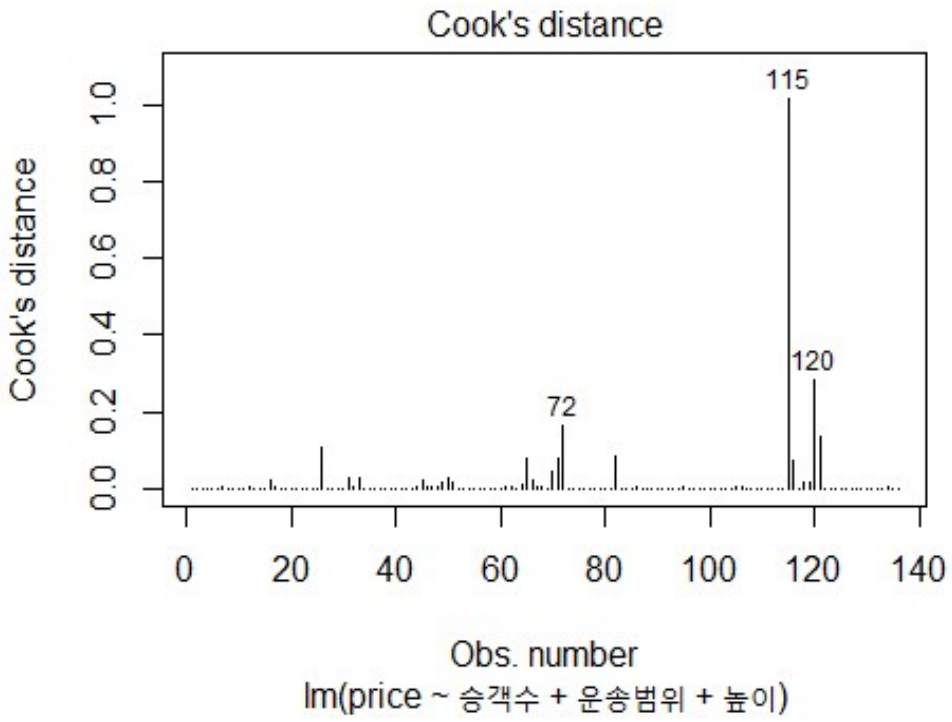
Scale-Location 의 그래프 결과 빨간색 실선이 수평선을 그리므로 등분산성을 만족한다고 판단하였다.

```
plot(fit.both,which=3)
```



3.2.5 이상치 확인

```
plot(fit.both,which=4)
```



115 번 데이터는 cook's distance 값이 1 보다 큰 이상치로 판단

나머지 데이터는 모두 0.4 미만이기 때문에 이상치로 판단하지 않음.

3.3 회귀모형 재 적합

115 번이상치를 제거하고 다시 회귀모형을 재적합한다.

```
fit_final=lm(price ~ 승객수 + 운송범위 + 높이, data = wf7[-c(115),])
summary(fit_final)
```

```
##
## Call:
## lm(formula = price ~ 승객수 + 운송범위 + 높이, data = wf7[-c(115),
##     ])
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -174.186  -10.236    3.639   13.158  167.466
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.386460   9.073179  -0.153   0.8788
## 승객수       0.445362   0.049137   9.064 1.57e-15 ***
## 운송범위     0.009097   0.001464   6.211 6.46e-09 ***
## 높이        -3.794324   1.668106  -2.275   0.0246 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.99 on 131 degrees of freedom
## Multiple R-squared:  0.8132,  Adjusted R-squared:  0.8089
## F-statistic: 190.1 on 3 and 131 DF,  p-value: < 2.2e-16
```

최종회귀모형식

$$\hat{y} = -1.386469 + 0.445362x_1 + 0.009097\,x_2 - 3.794324\,x_3$$

제 4 장 시각화

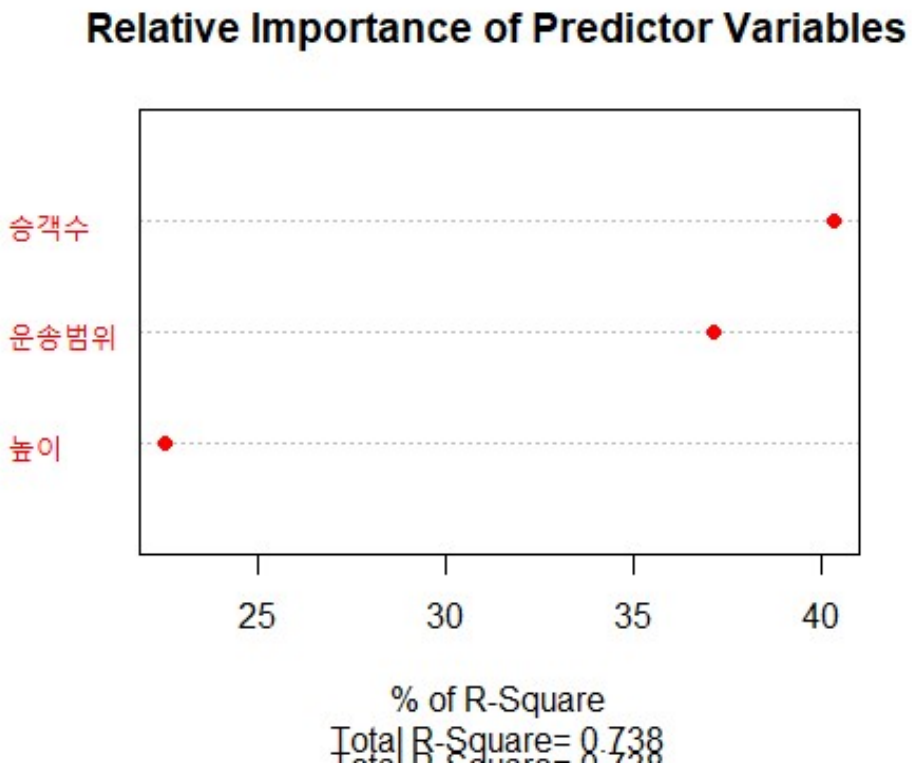
결과 예측에 가장 중요한 변수를 서열화하여 시각화하였다. (코드는 Dr.Johnson 논문에 소개된 변수의 상대적 중요도를 알 수 있는 스크립트를 인용하였다.)

그 결과 1 위는 승객수로 41.6%, 2 위는 운송 범위로 32.8%, 3 위는 높이로 25.6%를 차지하였다.

```
model=lm(price ~ 승객수 + 운송범위 + 높이, data = wf7[-c(115),])
relweights <- function(fit,...){
  R <- cor(fit$model)
  nvar <- ncol(R)
  rxx <- R[2:nvar, 2:nvar]
  rxy <- R[2:nvar, 1]
  svd <- eigen(rxx)
  evec <- svd$vectors
  ev <- svd$values
```

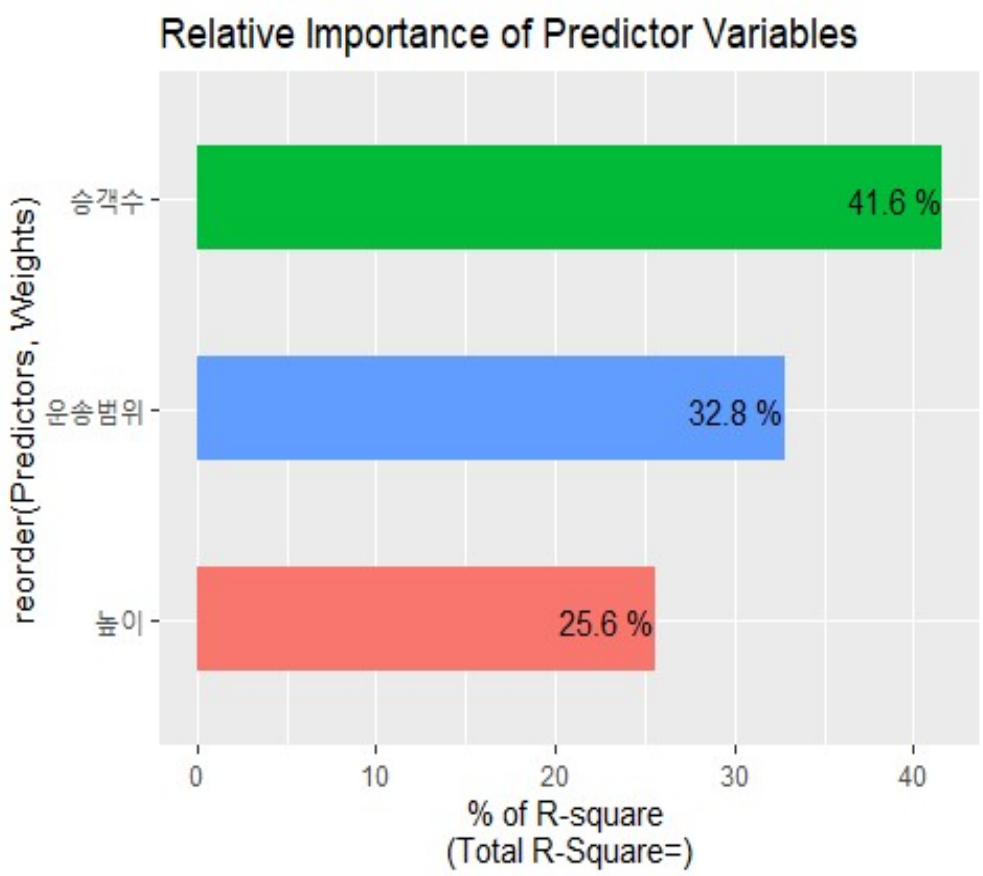
```
  delta <- diag(sqrt(ev))
  lambda <- evec %*% delta %*% t(evec)
  lambdasq <- lambda ^ 2
  beta <- solve(lambda) %*% rxy
  rsquare <- colSums(beta ^ 2)
  rawwgt <- lambdasq %*% beta ^ 2
  import <- (rawwgt / rsquare) * 100
  import <- as.data.frame(import)
  row.names(import) <- names(fit$model[2:nvar])
  names(import) <- "Weights"
  import <- import[order(import),1, drop=FALSE]
  dotchart(import$Weights, labels=row.names(import),
           xlab="% of R-Square", pch=19,
           main="Relative Importance of Predictor Variables",
           sub=paste("Total R-Square=", round(rsquare, digits=3)),
           ...)
  return(import)}
```

```
#변수의 상대적 중요도를 시각화
model_final = lm(price ~ 승객수 + 운송범위 + 높이, data = wf7)
result = relweights(model_final, col='red')
```



ggplot2 을 사용하여, 변수의 상대적 중요도를 시각화

```
library(ggplot2)
plotRelWeights=function(fit){
  data<-relweights(fit)
  data$Predictors<-rownames(data)
  p<-ggplot(data=data,aes(x=reorder(Predictors,Weights),y=Weights,fill=Predic
tors))+
    geom_bar(stat="identity",width=0.5)+
    ggtitle("Relative Importance of Predictor Variables")+
    ylab(paste0("% of R-square \n(Total R-Square=",attr(data,"R-square"),")
"))+
    geom_text(aes(y=Weights-0.1,label=paste(round(Weights,1),"%")),hjust=1)+
    guides(fill=FALSE)+
    coord_flip()
  p
}

model_3 = lm(price ~ 승객수 + 운송범위 + 높이, data = wf7[-c(115),])

plotRelWeights(model_3)
```



제 5 장 결과 요약

① 회귀직선이 전체 종속변수 값의 변화 중 약 80.9%를 설명함

② 승객수를 제외한 독립변수가 고정되어 있을 때,

승객수가 1 명 증가할 때 비행기 가격은 0.445362(백만 달러)만큼 증가

③ 운송범위를 제외한 독립변수가 고정되어 있을 때,

운송범위가 1km 증가할 때 비행기 가격은 0.009097(백만 달러) 만큼 증가

④ 높이를 제외한 독립변수가 고정되어 있을 때,

높이가 1m 증가할 때 비행기 가격은 3.794324(백만 달러) 만큼 감소