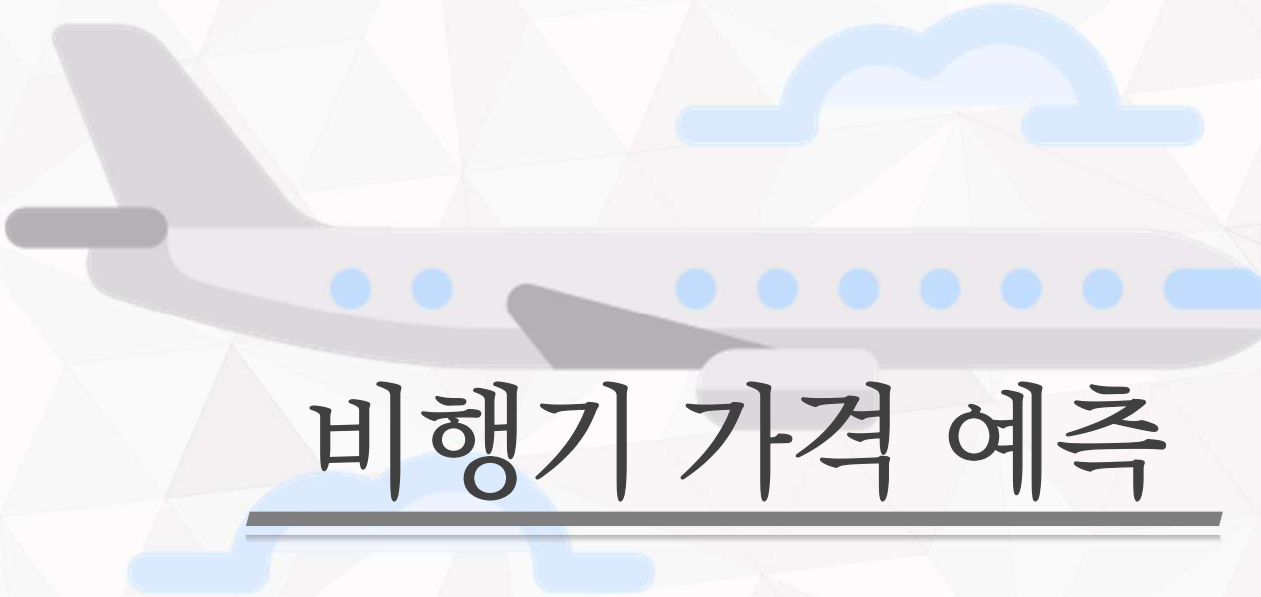
다중회귀분석

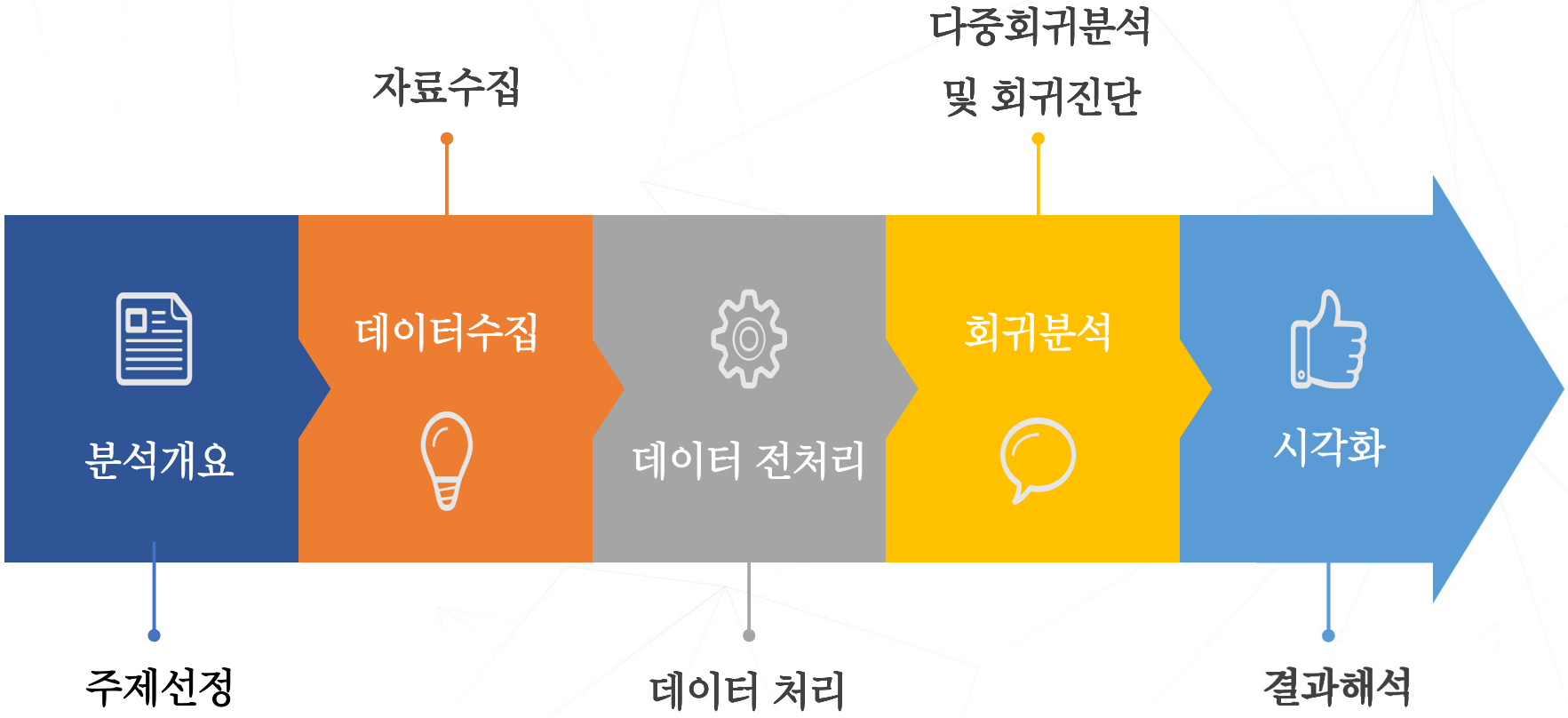
을 이용한



비행기 가격 예측

2019 정다인
2019 정솔잎
2018 박주연
20180769 박혜연
2018 이가은

INTRODUCTION



주제 선정

자동차 가격도 엔진, 제조사, 크기 등에 의해 결정이 되는데,
비행기의 가격은 어떤 변수에 의해 결정이 될까??

데이터 수집

DATA 공공데이터포털 .GO.KR

CSV **국토교통부_세계항공기_정보**

세계 항공기 기종별 제작사별 정보 제공 (제조사,모델,비행기구분,최초운항일 등)

	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	CONVAIF	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 HAVII	DHC-2 Turb	DH2T	DHR	LandPlane	LOW	STOLutility	Turboprop	1	1947-08-16	1657		Produc	1	6	9.2	2.7	732
7	DE HAVII	DHC-2 Beav	DHC2	DHP	LandPlane	LOW	STOLutility	Piston	1	1947-08-16	1657		Produc	1	6	9.2	2.7	732
8	DE HAVII	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 HAVII	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 HAVII	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 C	G159	GRS	LandPlane	MEDIUM	Business a	Turboprop	2	1958-08-14	200		Produc	2	24	19.4	6.9	4090
12	SIKORSK	Sikorsky S-5	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

296 * 18

데이터 처리

데이터 불러오기 및 확인하기

```
worldflights=read.csv("worldflights.csv",header=T)
head(worldflights)
```

```
             제조사                          비행기모델 ICAO.CODE IATA.CODE 비행기구분 비행기크기
           CURTISS              Curtiss C-46 Commando       C46       CWC  LandPlane     MEDIUM
           CONVAIR           Convair CV-240 / CV-440      CVLP       CV4  LandPlane     MEDIUM
         LOCKHEED Lockheed L-1049 Super Constellation      CONI       L49  LandPlane     MEDIUM
           DOUGLAS                     Douglas DC-6       DC6       D6F  LandPlane     MEDIUM
DE HAVILLAND CANADA              DHC-2 Turbo Beaver      DH2T       DHR  LandPlane        LOW
DE HAVILLAND CANADA                   DHC-2 Beaver      DHC2       DHP  LandPlane        LOW
                     CLASS          엔진타입 엔진수 최초운항일 생산수량 단가.백만달러.
Military transport aircraft            Piston      2 1940-03-26     3181
                   Airliner            Piston      2 1947-03-16     1181
                   Airliner            Piston      4 1951-07-14      259
Airliner/transport aircraft            Piston      4 1946-02-15      704
      STOLutility transport Turboprop/Turboshaft      1 1947-08-16     1657
      STOLutility transport            Piston      1 1947-08-16     1657
                       상태 조종석수 승객수 길이.m. 높이.m. 운송범위.km.
Production completed (1945)         5     40    23.3     6.6          870
Production completed (1954)         3     40    22.8     8.2         1900
Production completed (1958)         5    106    34.6     7.5         8288
Production completed (1958)         4     68    30.7     8.7         7377
Production completed (1967)         1      6     9.2     2.7          732
Production completed (1967)         1      6     9.2     2.7          732
```

결측치 존재

데이터 처리

변수 이름 바꾸기

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

종속변수(price)와 독립변수 12개 선택

```
library(dplyr)
```

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

데이터 처리

가격(price) 변수 숫자형으로 변환

```
> is.numeric(wf2$price)
[1] FALSE
> wf2$price=as.numeric(wf2$price)
> is.numeric(wf2$price)
[1] TRUE
```

```
> wf2$price
  [1]    NA    NA    NA    NA    NA    NA   0.1   0.1    NA    NA    NA    NA    NA
 [18]   0.2    NA   5.2    NA    NA    NA    NA    NA    NA   4.9  24.5   1.0    NA
 [35]    NA    NA    NA    NA    NA   0.2    NA    NA    NA   9.8 147.5    NA    NA
 [52]    NA  21.1    NA 146.7   0.5   0.1   4.3   1.5    NA   3.7   3.4   2.0  23.4
 [69]  20.0    NA    NA    NA   6.0   5.2    NA   4.0    NA    NA  48.5    NA    NA
 [86]  20.0  30.0 101.5 110.6 129.5   0.1    NA    NA  77.4  89.6 101.0 115.0 238.5
```

데이터 처리

범주형 데이터 처리

제조사, 비행기크기, 비행기 구분, 엔진타입은
사칙 연산 관계가 존재하는 것은 아니기 때문에
숫자로 취급하기보다는
범주형 변수인 factor로 처리

```
wf7$제조사 = factor(wf7$제조사)
wf7$비행기크기 = factor(wf7$비행기크기)
wf7$비행기구분= factor(wf7$비행기구분)
wf7$엔진타입 = factor(wf7$엔진타입)
```

결측치 제거

결측치가 포함되면 회귀분석 진행이 어려움.
가격, 승객수, 길이, 높이, 운송범위, 생산수량
결측치가 포함된 행은 모두 제거

```
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)
```

Sample size :136

```
> dim(wf7)
[1] 136  18
```

회귀 모형 적합

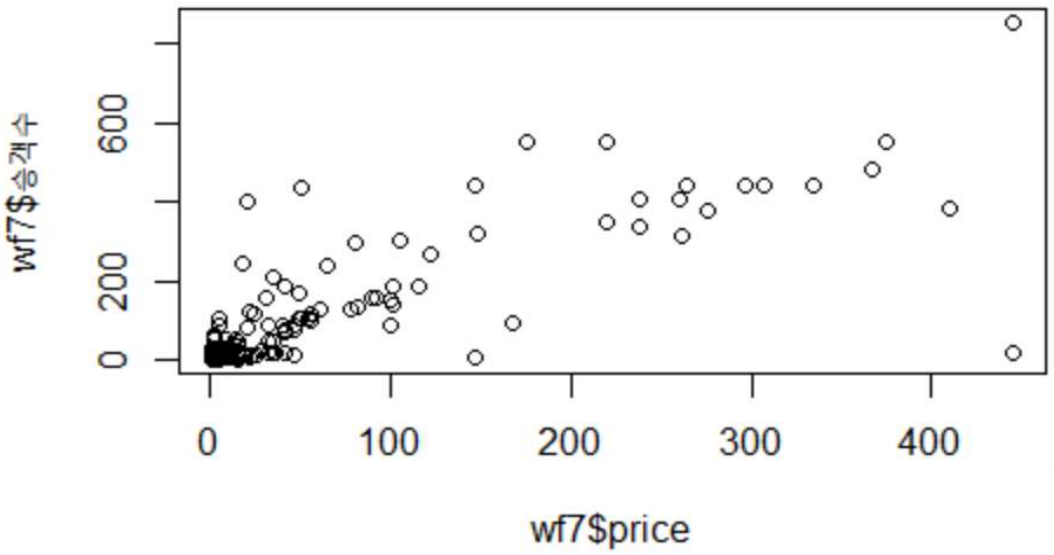
데이터 특성 파악

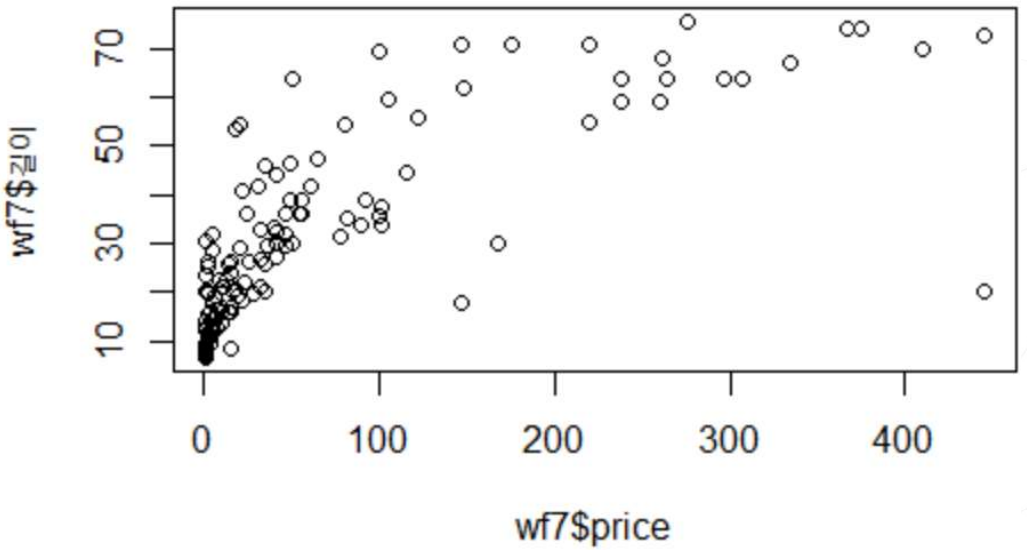
library(psych)

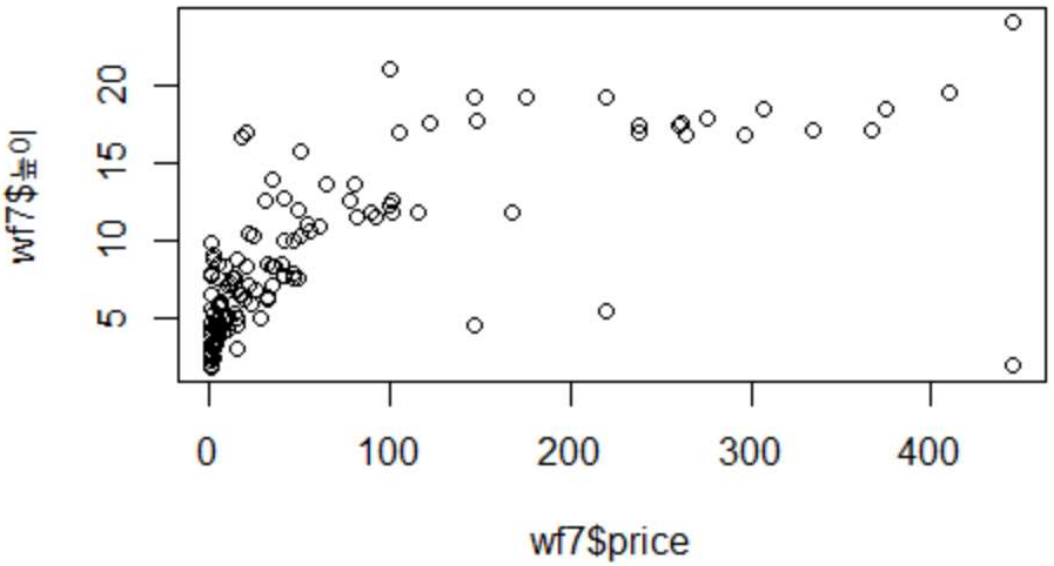
```
pairs.panels(wf7)
```

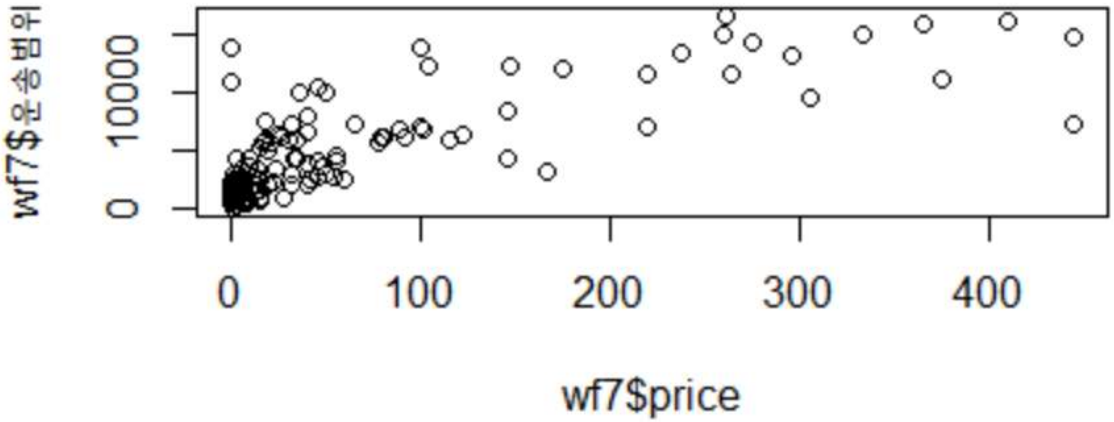
가격과 상관계수가 높은 변수
: 승객수, 길이, 높이, 운송범위

가격과 상관계수가 높은 4가지 변수의 산점도









회귀 모형 적합

회귀 분석 수행

여러 개의 독립변수를 가지므로 다중회귀분석을 실행

- **종속변수**
  가격(price)
- **독립변수**
  제조사, 비행기구분, 비행기 크기, class, 엔진타입, 엔진 수, 생산수량, 조종석 수, 승객 수, 길이, 높이, 운송범위

가설 세우기

H_0: $b_1 = b_2 = \cdots = b_{12} = 0$ (모든 독립변수는 유의한 영향을 미치지 않는다)

모든 독립변수는 price에 영향을 미치지 않을 것이다.

H_1: H_0 is not true (독립변수 중 적어도 하나는 유의한 영향을 미친다)

price에 적어도 하나는 영향을 미칠 것이다.

회귀 모형 적합

- R code

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

- F-test결과

```
---
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
```

P-value가 매우 작기 때문에 귀무가설 기각
즉, 적어도 하나의 독립변수는 종속변수에 유의한 영향을 미친다

회귀 모형 적합

변수 선택 1. 전진선택법 (forward)

변수 추가 시마다 p값이 낮은 유의한 변수를 하나씩 추가함

R code

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

summary(fit.forward1)
```

결과

```
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
```

유의한 변수
: 승객 수, 운송 범위, 높이

결정계수
: 약 73.2%

회귀 모형 적합

변수 선택 2. 후진제거법 (backward)

유의하지 않은 변수가 많아도 p값을 기준으로 전체에서
1개씩 제거하여 모든 변수가 유의할 때 까지

R code

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

결과

```
lm(formula = price ~ 제조사 + class + 엔진수 + 생산수량 + 길이 +
    높이 + 운송범위, data = wf7)

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
```

회귀 모형 적합

변수 선택 3. 단계선택법(stepwise)

모든 부분집합을 고려하는 방법으로 best 변수를 선택함

- R code

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

summary(fit.both)
```

- 결과

```
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
```

Forward와 같은 결과
이 모형 선택

회귀계수 진단

- 가설 세우기

  H_0: $b_i = 0$ (i=1~3) (회귀계수는 유의하지 않다)

  H_1: H_0 is not true (회귀계수는 유의한 영향을 미친다)

- T-test결과

```
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
```

회귀계수의 유의성 확인
귀무가설 기각
모든 회귀계수가 유의한 결과

회귀모형 진단

- 회귀 모형의 설명력

$$\hat{y} = 12.692479 + 0.470124x_1 + 0.011942x_2 - 7.019925x_3$$

```
Multiple R-squared:  0.738,    Adjusted R-squared:  0.7321
```

조정된 결정계수 확인하기
회귀 직선이 자료의 73.21%를 설명함

- 변수들 간의 다중공선성 확인

```
> vif(fit.both)
   승객수  운송범위      높이
4.665014 2.918100 5.583548
```

10 미만이므로 심각한 문제는 없다고 해석할 수 있다
즉, 독립변수들 사이에 심각한 선형관계는 없다

회귀모형 진단

선형회귀모형의 기본 가정 4가지 : 선형성, 독립성, 정규성, 등분산성

✓ **독립성** 오차항들 간의 자기상관 확인

: 더빈-왓슨 통계량으로 오차항의 자기상관성 여부 검정

```
> dwtest(fit.both)
```

library(lmtest)

```
        Durbin-Watson test

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

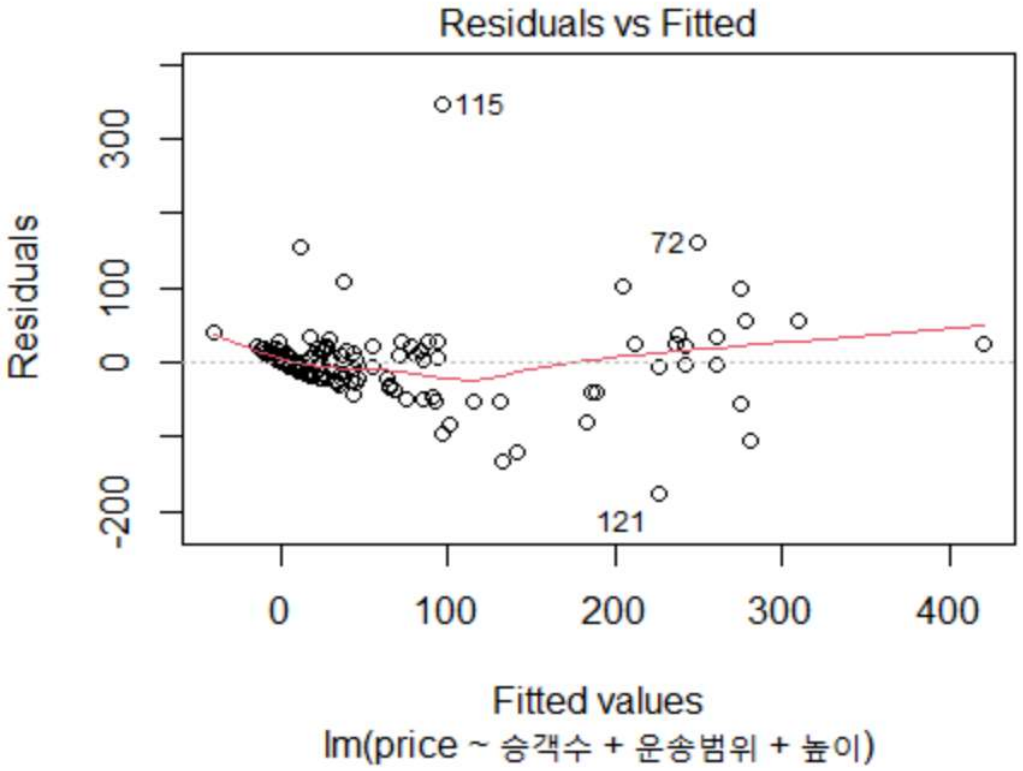
➡ 2에 가까운 값을 가지므로 자기상관에 문제가 없다고 판단

회귀모형 진단

선형회귀모형의 기본 가정 4가지

```
plot(fit.both)
```

✓ 선형성

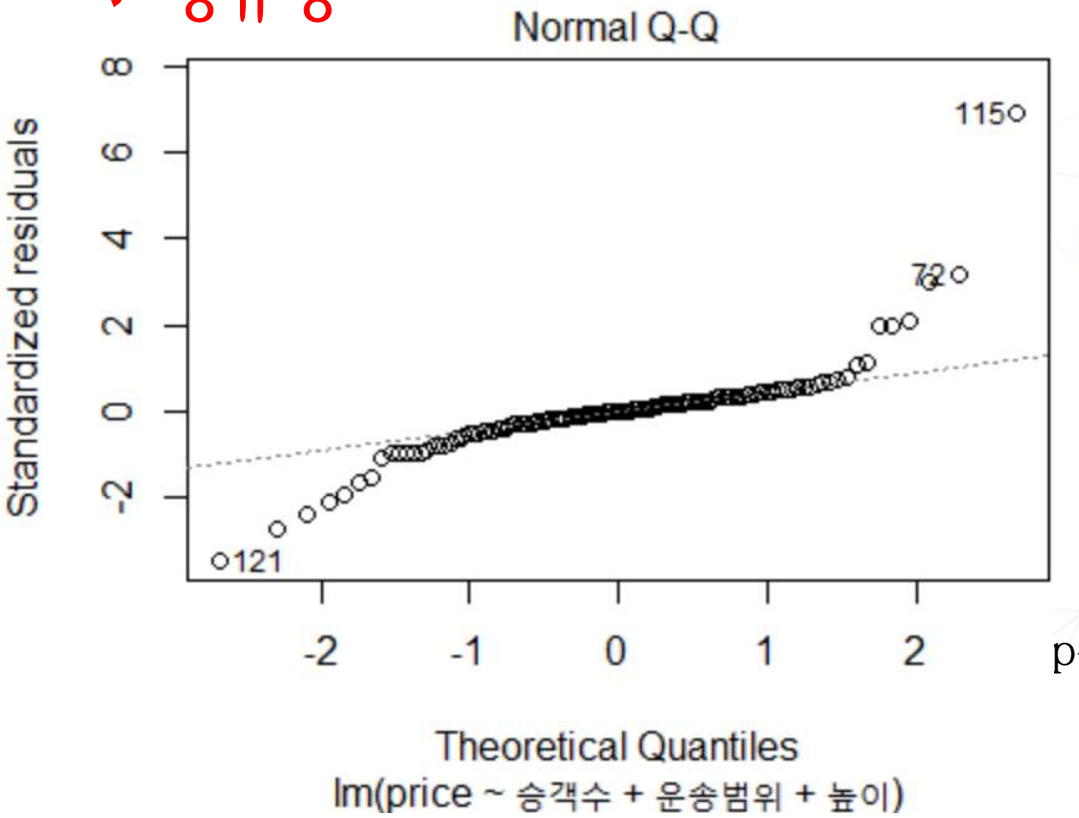


- 예측값(fitted)과 잔차(residual)의 비교
- 빨간 실선은 잔차의 추세를 나타냄
- 빨간 실선이 점선에서 크게 벗어나지 않으므로 예측값에 따라 잔차가 크게 달라지지 않는다 판단

회귀모형 진단

선형회귀모형의 기본 가정 4가지

✓정규성



샤피로의 검정으로 확인하기

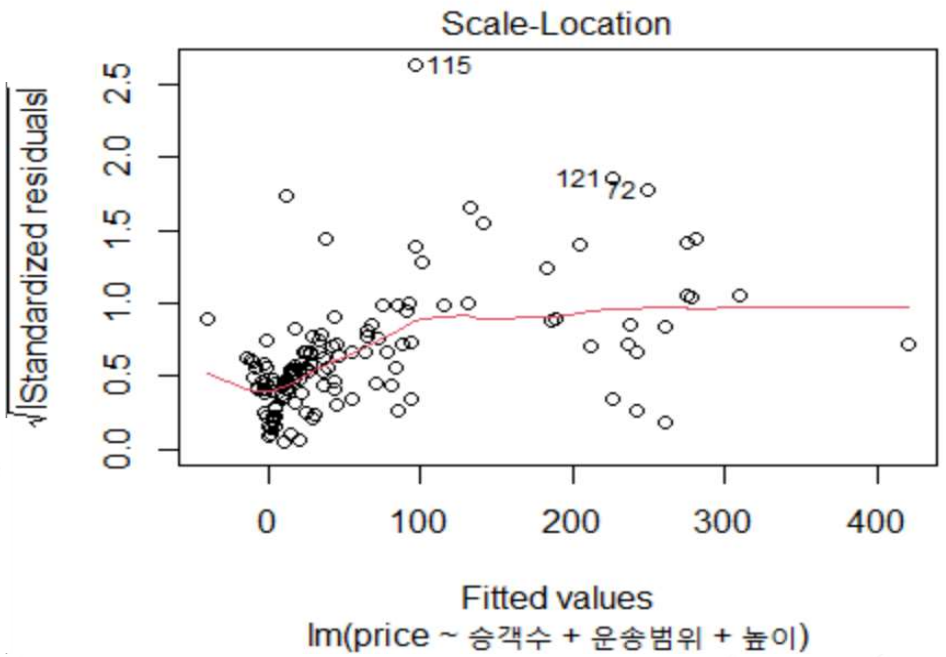
```
> shapiro.test(fit.both$residuals)

        Shapiro-Wilk normality test

data:  fit.both$residuals
W = 0.7593, p-value = 1.175e-13
```

p값이 매우 작으므로 잔차의 정규성이 위반되지 않는다 판단

회귀모형 진단

선형회귀모형의 기본 가정 4가지

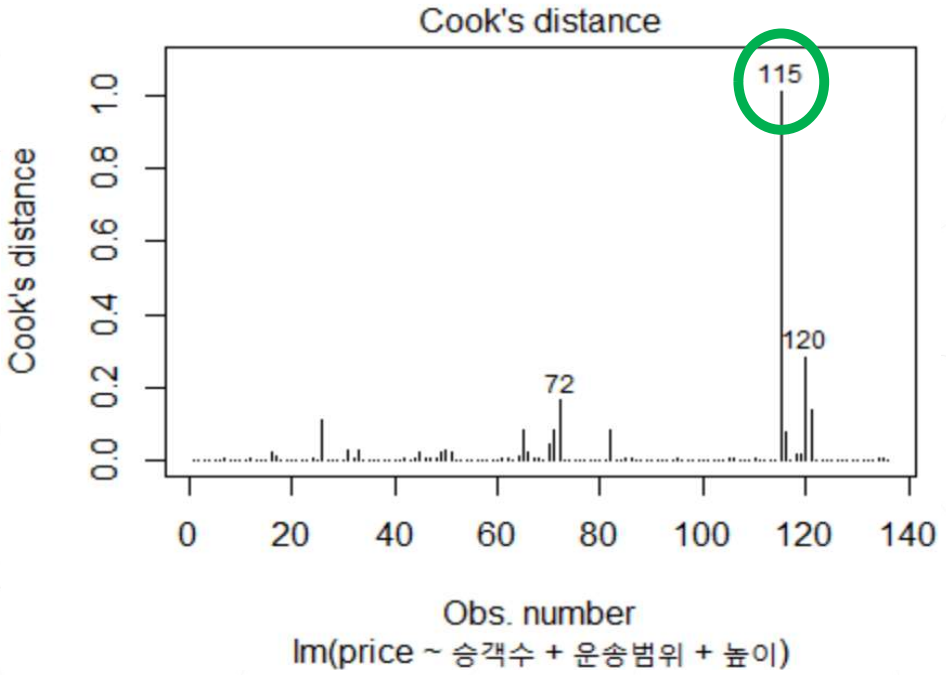
✓ 등분산성



빨간색 실선이 수평선을 그리는 것이 이상적

등분산성 만족한다 판단

✓ 이상치



Cook's distance 극단값을 나타내는 지표

회귀모형 재적합

이상치 제거

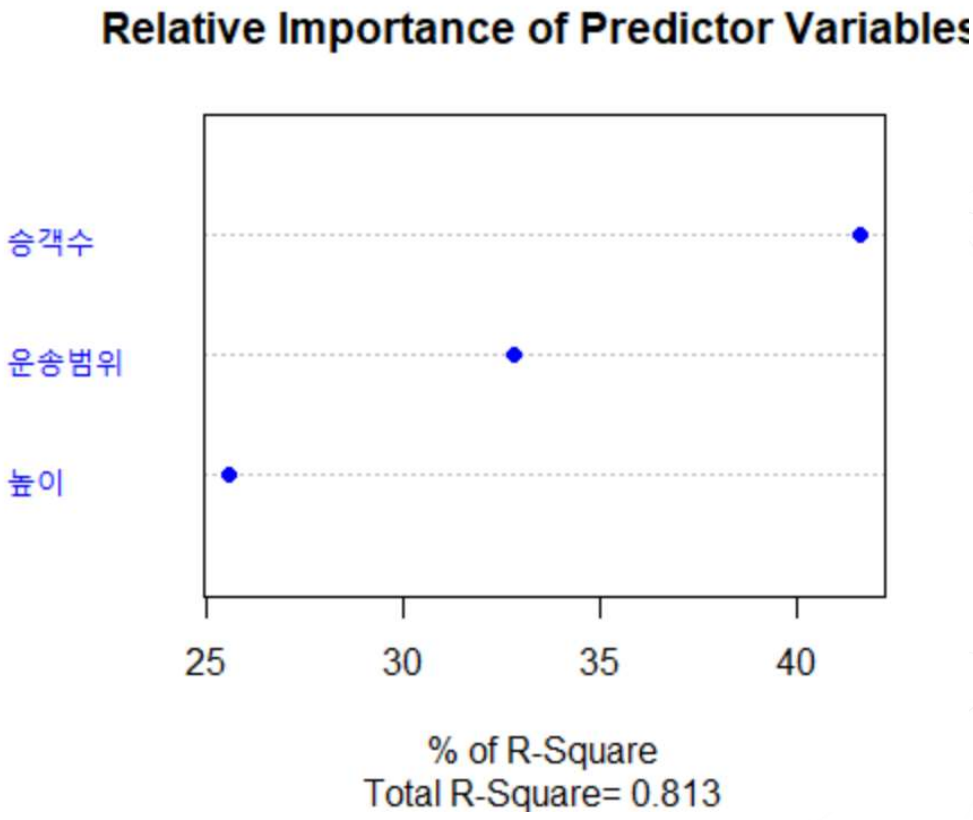
```
fit_final=lm(price ~ 승객수 + 운송범위 + 높이, data = wf7[-c(115),])
summary(fit_final)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.386460   9.073179  -0.153   0.8788
승객수       0.445362   0.049137   9.064 1.57e-15 ***
운송범위     0.009097   0.001464   6.211 6.46e-09 ***
높이        -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
```

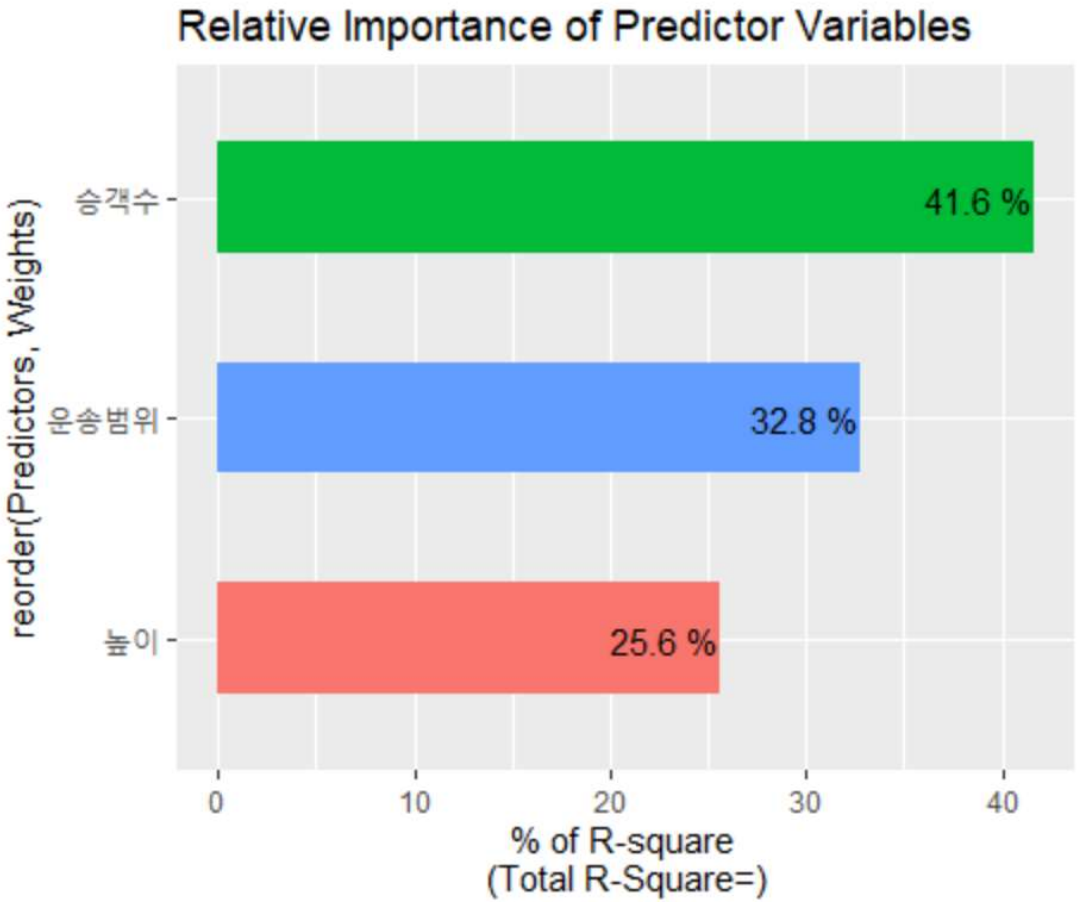
조정된 결정계수 증가
0.7321->0.8089

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

시각화

변수의 상대적 중요도를 시각화





결과 해석

비행기 가격에는 승객수, 운송범위, 높이가 영향을 미친다

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

x_1: 승객수, x_2 : 운송범위x_3 : 높이

① 회귀직선이 전체 종속변수 값의 변화 중 약 80.9%를 설명함

② 승객수를 제외한 독립변수가 고정되어 있을 때,
승객수가 1명 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


감사합니다