비타민 19.11.27.(수) Session

회귀진단 (Regression Diagnostics)

회귀모델의 적합성 판정+구간추정+다중공선성

2 조

이지선

우현우

Index

1. 회귀분석 결과 그래프 해설

2. 점 추정과 구간 추정

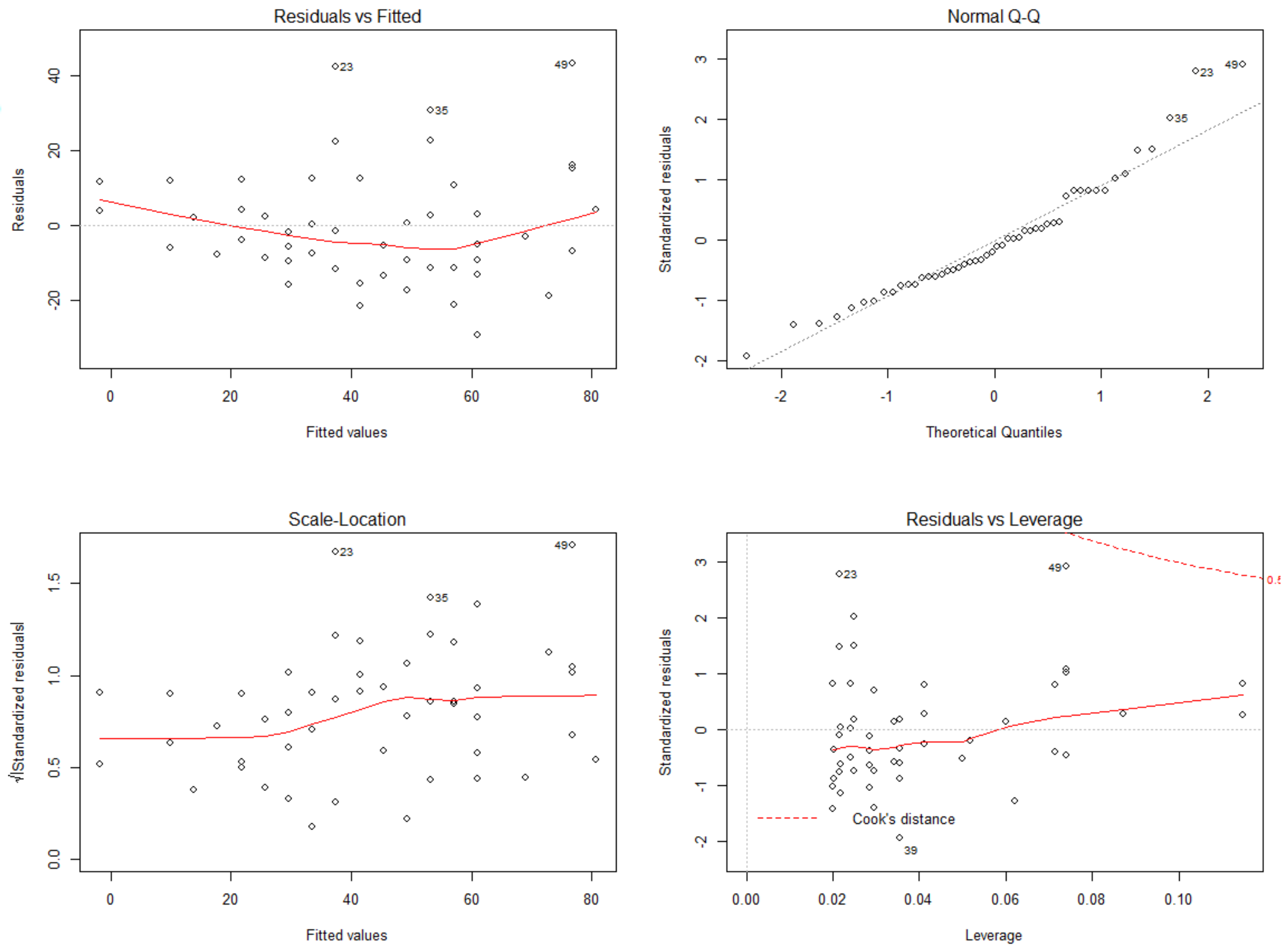
3. 실습

4. 독립변수 선정에서의 유의사항
 - “다중공선성”

1. 회귀분석 결과 그래프 해석

회귀모델의 적합성 판정

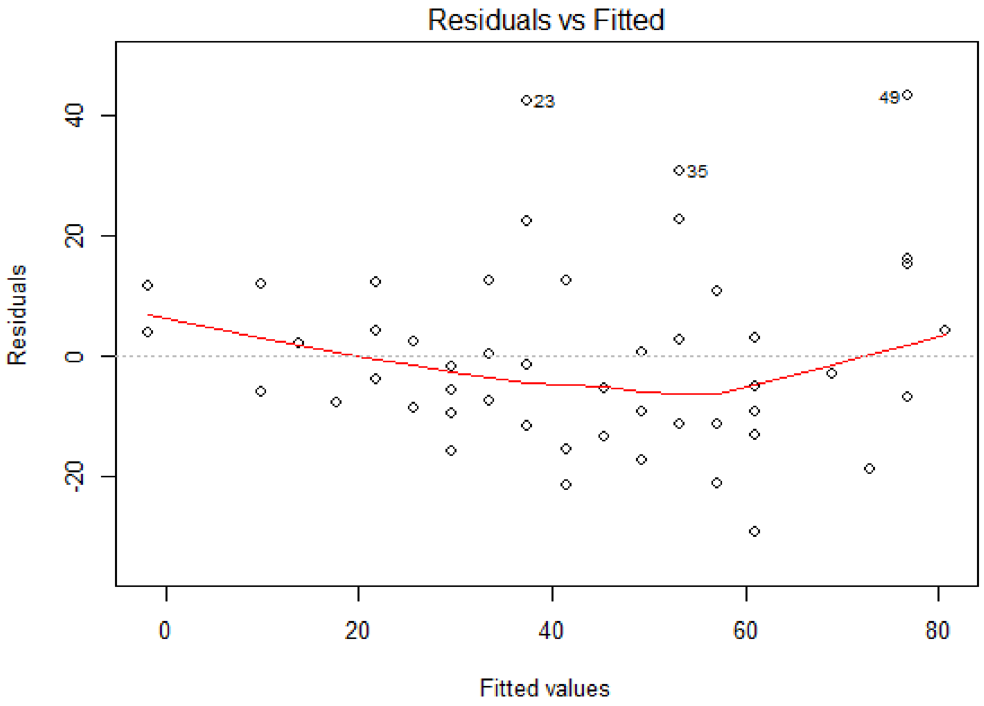
```
> lm_result <- lm(formula = dist~speed, data=cars)
> par(mfrow=c(2,2))
> plot(lm_result)
```



1. Residuals vs Fitted
2. Normal Q-Q
3. Scale-Location
4. Residuals vs Leverage

1. 회귀분석 결과 그래프 해석

1. Residuals vs Fitted



< Residuals vs Fitted >

X 축 : 선형 회귀로 예측된 Y 값
Y 축 : 잔차 (Residuals)

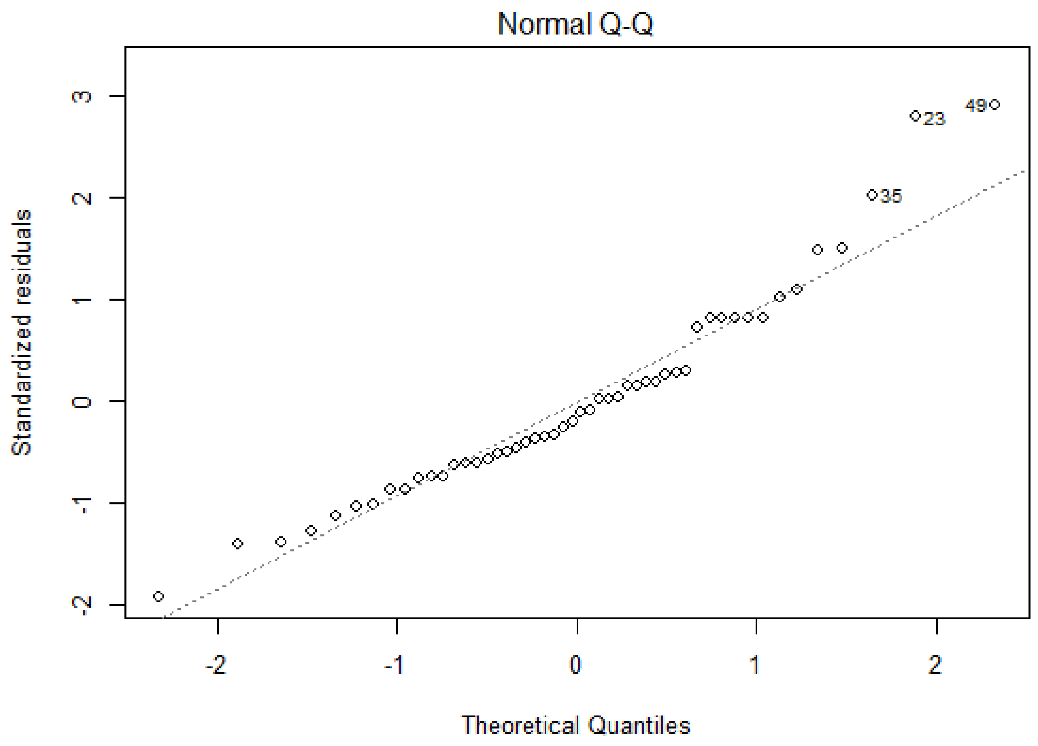
> 선형 회귀에서 오차는 평균이 0이고 분산이 일정한 정규 분포를 가정함.

> 따라서 예측된 Y값과 무관하게 잔차의 평균은 0이고 분산은 일정해야 함.

> 따라서 이 그래프에서는 기울기가 0에 가까운 직선이 관측되는 것이 이상적

1. 회귀분석 결과 그래프 해석

2. Normal Q-Q



< Normal Q-Q >

> 잔차가 정규 분포를 따르는지 확인하는 그래프

점 (x,y) : x = 첫 번째 분포의 첫 번째 quantile
y = 두 번째 분포의 첫 번째 quantile

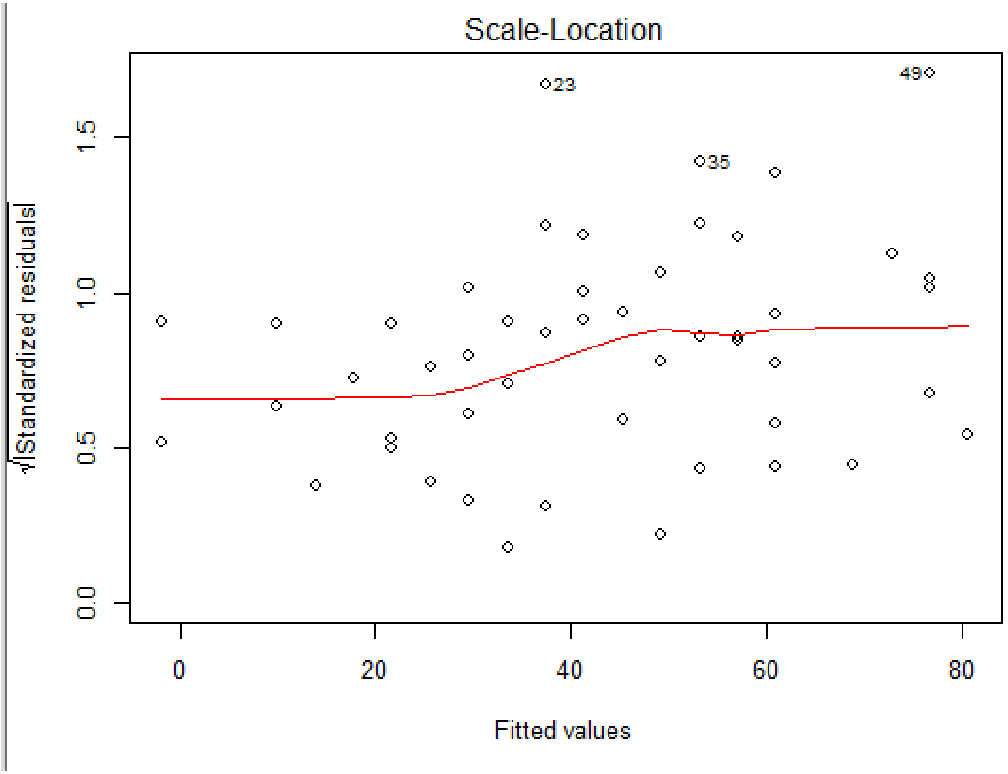
Ex) 100-quantiles 라면 두 분포에서 1%에 해당하는 샘플이 각각 x, y값이 되어 해당 점을 찍음.

> 만약 샘플 수가 다르다면 적은 샘플 수를 가진 샘플이 중복하여 퍼센트를 기준으로 점을 찍는다.

> 직선에 점들이 밀집되어 있을수록 이상적 (좋은 모델)

1. 회귀분석 결과 그래프 해석

3. Scale-Location



< Scale - Location >

X 축 : 선형 회귀로 예측된 Y 값
Y 축 : 표준화 잔차 (Standardized Residuals)

> 잔차 (실제 값과 예측 값의 차이)가 등분산성을 따르는지 확인하는 그래프

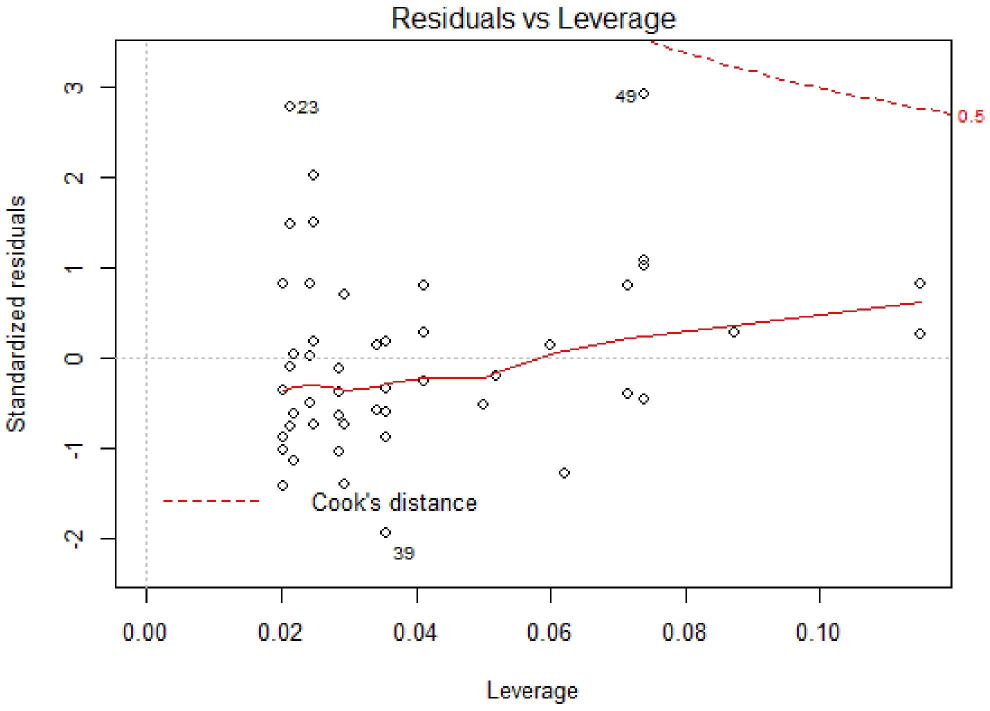
> 기울기가 0에 가까운 직선이 관측되는 것이 이상적

> 특정 위치에서 0에서 멀리 떨어진 값이 관찰된다면 해당 점에 대해 표준화 잔차가 크다, 즉 회귀 직선이 해당 Y값을 잘 적합하지 못함을 의미

→ 이상치(outlier)일 가능성이 있다

1. 회귀분석 결과 그래프 해석

4. Residuals vs Leverage



< Residuals vs Leverage >

X 축 : 레버리지
(Leverage, <u>독립 변수가 얼마나 극에 치우쳐 있는지를 의미</u>
특정 데이터만 값이 극단적으로 크거나 작으면 레버리지가 크다고 함.)

Y 축 : 잔차 (Residuals)

> 이상치를 표현하는 그래프

> 쿡의 거리 (Cook's Distance)
: 회귀 직선의 모양에 크게 영향을 미치는 점들을 찾는 방법
레버리지와 잔차에 비례함. 따라서 우측 상단과 우측 하단에 쿡의 거리가 큰 값들이 위치함.

> 빨간 점선 안에 점들이 들어있지 않을수록 좋은 모델

2. 점 추정과 구간 추정

새로운 독립변수를 회귀모델 방정식에 대입해 종속변수를 예측할 수 있다.

1. 점 추정 (Point Estimation) 2. 구간 추정 (Interval Estimation)

점 추정 (Point Estimation)

: **종속변수 값을 특정 값 하나로 예측**

장점 : 간단명료하게 표현 가능
("가장 좋은 단일 예측 값" 을 제시)

단점 : 예측 값의 불확실성을 표현하지 못함

< 불확실성 >

1. 회귀 모델 방정식의 계수에 대한 불확실성
2. 회귀 모델 방정식을 통해 나온 결과값에 대한 불확실성
   (결과 오차 범위)

2. 점 추정과 구간 추정

점 추정

$$E\{Y\} = \beta_0 + \beta_1 X$$

E{Y} (mean response) : X에 대한 Y의 확률분포의 평균

< 예측된 회귀 함수 식 >

$$\hat{Y} = b_0 + b_1 X$$

Yhat : 독립 변수 X에 대한 예측된 회귀 함수 값

= 독립 변수 X에 대한 E{Y}의 점 추정 (불편 추정량, 최소 분산)

$$\hat{Y}_i = b_0 + b_1 X_i \qquad i = 1, \ldots, n$$

Yihat : i번째 상황의 적합치

= i번째 X (Xi)에 대한 E{Y}의 점 추정 값

2. 점 추정과 구간 추정

새로운 독립변수를 회귀모델 방정식에 대입해 종속변수를 예측할 수 있다.

1. 점 추정 2. 구간 추정

구간 추정

: 점 추정의 불확실성을 감안해 종속변수를 하나의 값이 아닌 범위 값으로 제시하는 방식

< 표현 방식 >

"제동거리는 202 feet ~ 312 feet (신뢰구간) 사이일 확률이 95% (신뢰수준)다."

> 이 때 신뢰수준이 높고, 신뢰구간이 좁을수록 좋은 모델!

< 종류 >

1. 신뢰 구간 (Confidence Interval)

2. 예측 구간 (Predict Interval)

2. 점 추정과 구간 추정

1. 신뢰 구간 (Confidence Interval)

: 모델계수에 대한 불확실성을 감안한 구간 추정

= 각 X (독립변수)값에 대한 Y (종속변수)의 평균 범위 구간

> (오차의 평균을 0으로 가정하기 때문에 오차항 고려 x)

$$\hat{y}(x_0) = b_0 + b_1 x_0$$

$$\begin{aligned} \text{Var}\ \hat{y}(x_0) &= \text{Var}[b_0 + b_1 x_0] \\ &= \text{Var}[\bar{y} + b_1(x_0 - \bar{x})] \end{aligned}$$

추정 표준 오차
(estimated standard error)

$$s_{\hat{y}(x_0)} = s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

예측의 표준 오차
(standard error of prediction)

$$Var\ \hat{y}(x_0) = \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$$

$E(y|x_0)$ 에 대한 $100(1-\alpha)$ % 신뢰구간(confidence interval)

$$\hat{y}(x_0) \pm t_{1-\alpha/2,\, n-2}\, s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

1-(a/2)

2. 점 추정과 구간 추정

2. 예측 구간 (Predict Interval)

: 모델계수에 대한 불확실성과 결과의 오차까지 감안한 구간 추정

= 각 X (독립변수)값에 대해서 예측되는 단일 관측치 Y (종속변수)의 범위 구간

= x0에서 단일 관측치 y0가 실제로 속할 것으로 기대되는 범위를 반영

*** x=x0에서의 단일 관측치 y0는 yhat(x0)(=x0일 때 예측된 회귀값)과 무관함**

2. 점 추정과 구간 추정

2. 예측 구간 (Predict Interval)

$$Var(y_0 - \hat{y}(x_0)) = \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)$$

$$= \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)$$

양 (quantity)
= $y_0 - \hat{y}(x_0)$ 차이의 표준오차 (standard error)

$$s\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}$$

$$\frac{y_0 - \hat{y}(x_0)}{\sigma\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}} \sim N(0,1)$$

y_0 가 (1-)의 고정확률(fixed probability)로 포함되는 구간

$$\frac{y_0 - \hat{y}(x_0)}{s\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}} \sim t_{n-2}$$

$$\hat{y}(x_0) \pm t_{\alpha/2,\, n-2}\, s\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}$$

2. 점 추정과 구간 추정

신뢰구간 (Confidence Interval) vs 예측 구간 (Predict Interval)

예측 구간은 하나의 Y값에 대한 불확실성을 반영하고
신뢰 구간은 Y의 예측된 평균에 대한 불확실성을 반영한다.

> **따라서 예측 구간이 신뢰 구간보다 주로 큰 범위를 갖는다.**

1. 회귀 함수 (Linear Regression Line) : 파란색
2. 신뢰 구간 (Confidence Interval) : 회색
3. 예측 구간 (Predict Interval) : 빨간색

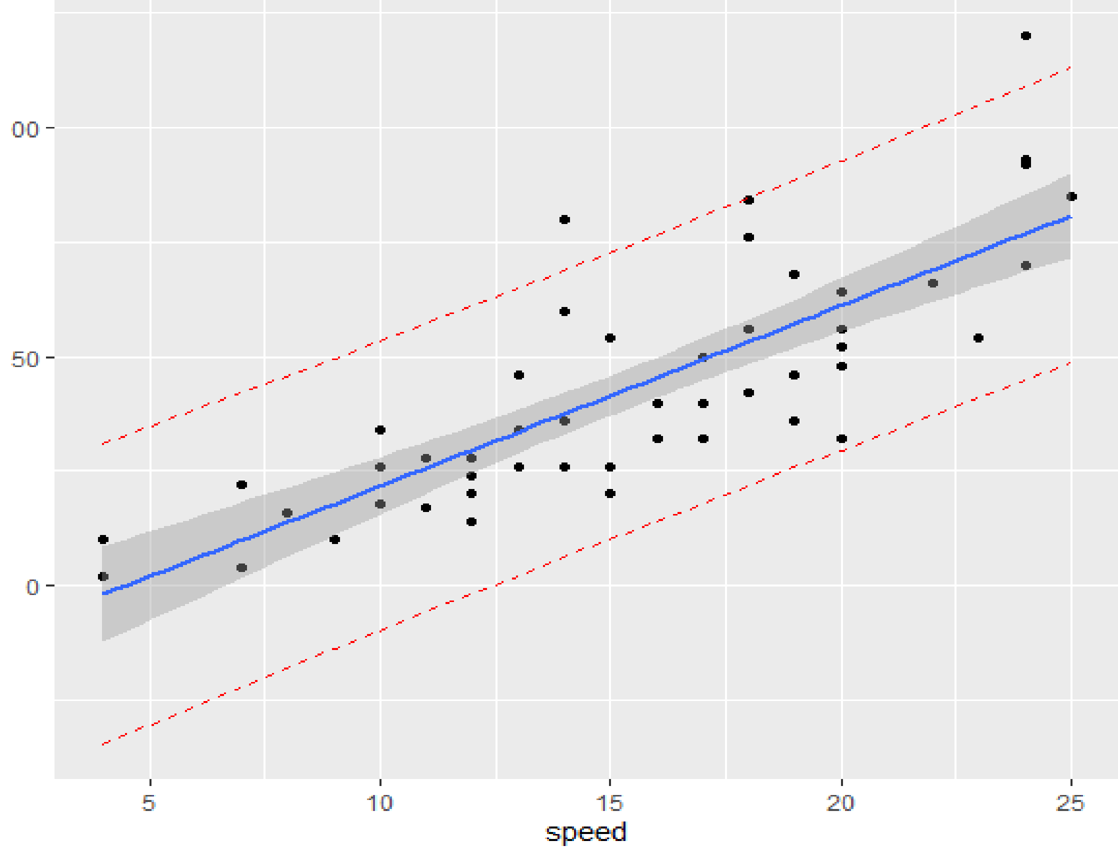
> 코딩 설명은 뒤에서 하겠습니다!

```r
# Build linear model
data("cars", package = "datasets")
model <- lm(dist~speed, data=cars)

# 1. Add predictions
pred.int <- predict(model, interval = "prediction")
mydata <- cbind(cars, pred.int)

# 2. Regression line + confidence intervals(신뢰구간)
library(ggplot2)
p <- ggplot(mydata, aes(speed,dist)) +
  geom_point()+
  stat_smooth(method = lm)

# 3. Add prediction intervals(예측구간)
p + geom_line(aes(y=lwr), color="red", linetype="dashed")+
  geom_line(aes(y=upr), color="red", linetype="dashed")
```



2. 점 추정과 구간 추정

Predict 함수

: **predict(object, data, interval, level)**

< 입력 항목 >

object : lm 함수를 통해 나온 회귀분석 결과

> 회귀 함수 식(을 담고 있는 변수)

data : 예측하고자 하는 독립변수를 담은 데이터프레임

*** 데이터프레임의 변수명(칼럼명)은 독립변수와 같아야 함**

interval : 구간 추정 시 사용하는 옵션 (default : "none")

> "none" : 점 추정 (point estimation) → default
> "confidence" : 신뢰 구간 추정 (confidence interval)
> "prediction" : 예측 구간 추정 (prediction interval)

level : 신뢰수준 (default : 0.95)

3. 실습

점 추정 (Point Estimation) 실습 1

```
# 점추정 (Point Estimation) 실습 1

# 차속도(speed)에 따른 제동거리(dist) 회귀분석
lm_result <- lm(formula=dist~speed, data=cars)

# 예측할 독립변수 데이터프레임 생성
# 데이터프레임을 생성할 때는 회귀분석 시 사용한 독립변수명과
# 동일하게 칼럼명 생성

speed <- c(50,60,70,80,90,100)
df_input <- data.frame(speed)

# 예측 - 점 추정 방식(point estimation)
# interval = "none" (interval 옵션 자체를 생략해도 됨)

predict(lm_result,df_input)
```

결과 :

```
> predict(lm_result,df_input)
       1        2        3        4        5        6 
179.0413 218.3654 257.6895 297.0136 336.3377 375.6618 
```

3. 실습

점 추정 (Point Estimation) 실습 1

```
# 점추정 가독성 고려하여 표현하기
predict_dist <- predict(lm_result, df_input)

# cbind를 사용하여 두 개의 데이터프레임
# df_input(새로운 speed 독립변수 값)과
# predict_dist(점 추정 결과값)을 가로로 연결
cbind(df_input, predict_dist)
```

결과 :

```
> cbind(df_input, predict_dist)
  speed predict_dist
1    50     179.0413
2    60     218.3654
3    70     257.6895
4    80     297.0136
5    90     336.3377
6   100     375.6618
```

3. 실습

신뢰 구간 추정 (Confidence Interval Estimation) 실습 2

```
# 신뢰구간 추정 (Confidence Interval Estimation) 실습 2

# 신뢰구간 추정 (interval = "confidence"), 신뢰수준 : 0.95
# 신뢰수준이 0.95이므로 level 생략 가능
# (참고) : lm_result = 차속도에 따른 제동거리 회귀분석
predict_dist <-
  predict(lm_result, df_input, interval="confidence")

# cbind 사용해서 df_input과 predict_dist 함께 보기
cbind(df_input, predict_dist)
```

결과 :

```
> cbind(df_input, predict_dist)
  speed      fit      lwr      upr
1    50 179.0413 149.8060 208.2766
2    60 218.3654 180.8489 255.8820
3    70 257.6895 211.8651 303.5139
4    80 297.0136 242.8670 351.1602
5    90 336.3377 273.8603 398.8151
6   100 375.6618 304.8480 446.4755
```

3. 실습

예측 구간 추정 (Prediction Interval Estimation) 실습 3

```r
# 예측구간 추정 (Prediction Interval Estimation) 실습 3

# 예측구간 추정 (interval = "prediction"), 신뢰수준 : 0.95
# 신뢰수준이 0.95이므로 level 생략 가능
# (참고) : lm_result = 차속도에 따른 제동거리 회귀분석
predict_dist <-
  predict(lm_result, df_input, interval="prediction")

# cbind 사용해서 df_input과 predict_dist 함께 보기
cbind(df_input, predict_dist)
```

결과 :

```
> cbind(df_input, predict_dist)
  speed      fit      lwr      upr
1    50 179.0413 136.4865 221.5962
2    60 218.3654 169.7474 266.9834
3    70 257.6895 202.4076 312.9715
4    80 297.0136 234.6592 359.3680
5    90 336.3377 266.6266 406.0488
6   100 375.6618 298.3908 452.9328
```

3. 점 추정과 구간 추정

신뢰구간 (Confidence Interval) vs 예측 구간 (Predict Interval)

```
# Build linear model
data("cars", package = "datasets")
model <- lm(dist~speed, data=cars)

# 1. Add predictions
pred.int <- predict(model, interval = "prediction")
mydata <- cbind(cars, pred.int)

# 2. Regression line + confidence intervals(신뢰구간)
library(ggplot2)
p <- ggplot(mydata, aes(speed,dist)) +
  geom_point()+
  stat_smooth(method = lm)

# 3. Add prediction intervals(예측구간)
p + geom_line(aes(y=lwr), color="red", linetype="dashed")+
  geom_line(aes(y=upr), color="red", linetype="dashed")
```

차속도에 따른 제동거리 회귀분석 (model) 생성

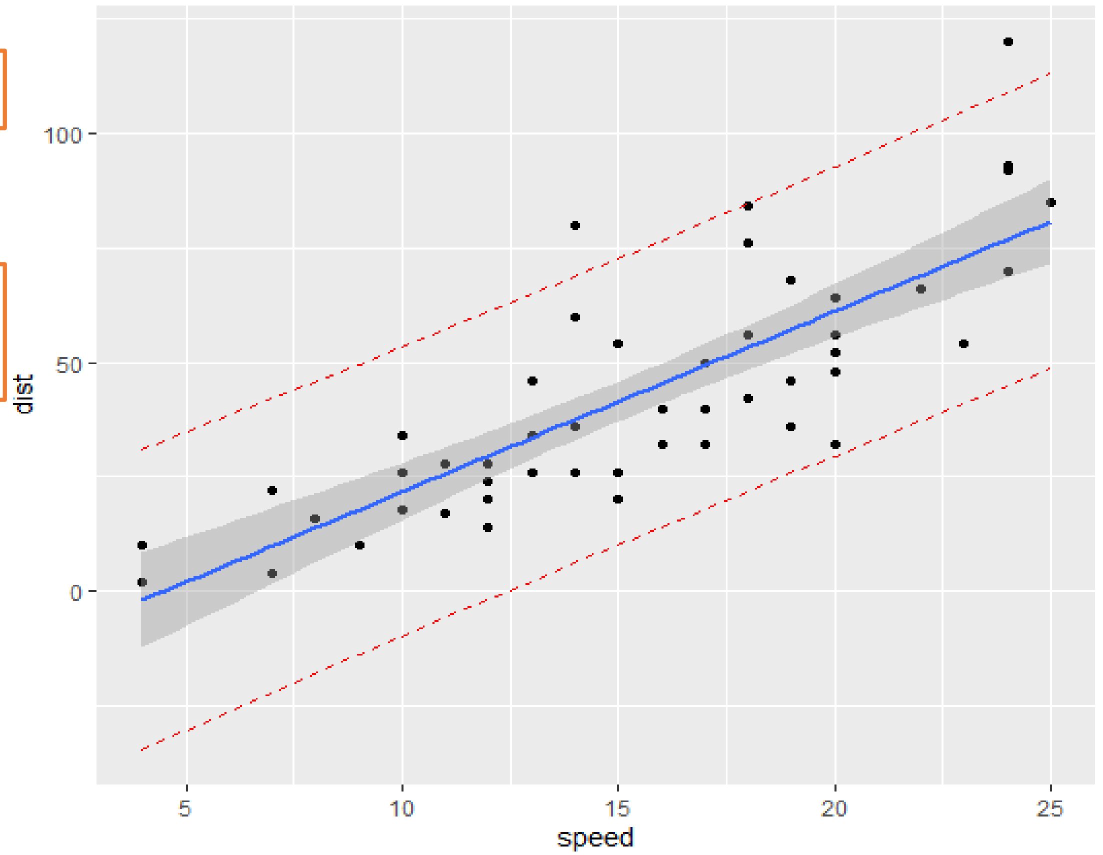
예측 구간 추정 실행

회귀 직선과 신뢰구간을 동시에 그려주는 코드

lwr : 구간 최솟값
upr : 구간 최댓값

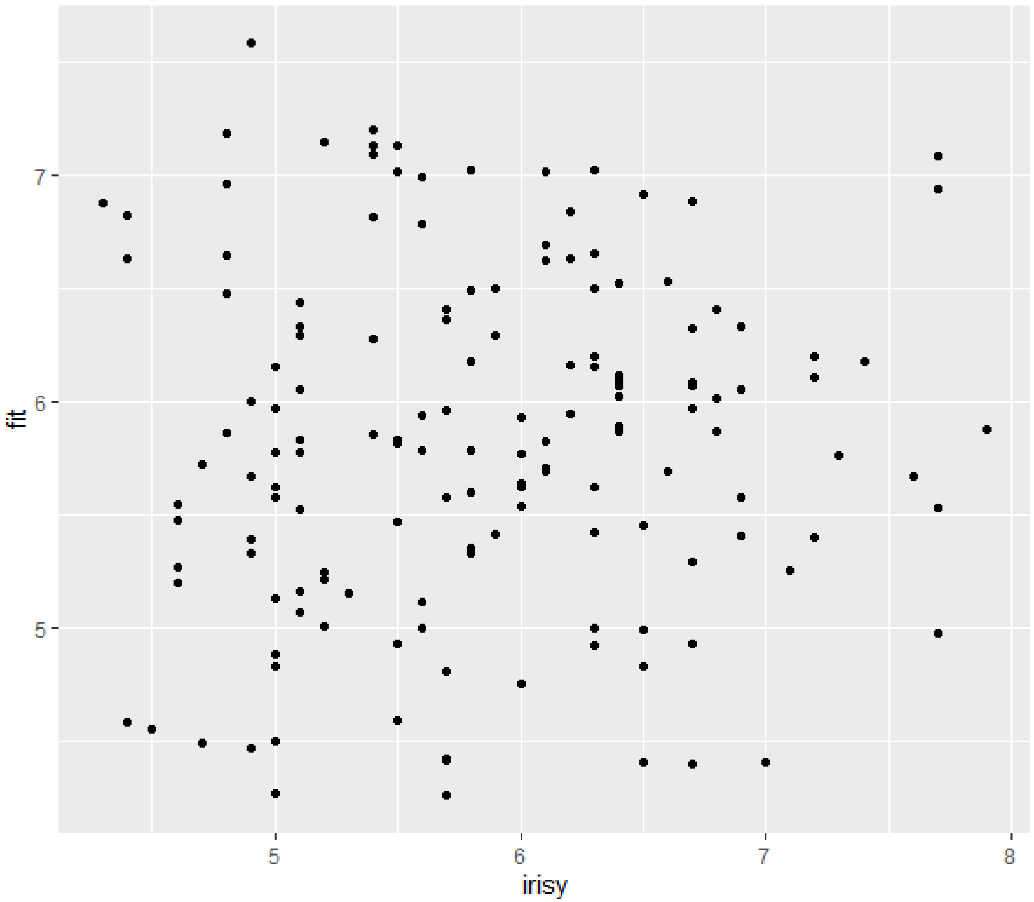
> 구간 최솟값을 하나의 점선으로 표현
> 구간 최댓값을 하나의 점선으로 표현

1. 회귀 직선 (Linear Regression Line) : 파란색
2. 신뢰 구간 (Confidence Interval) : 회색
3. 예측 구간 (Predict Interval) : 빨간색



+ y vs predict

```
str(iris)
iris_lm <- lm(Sepal.Length ~ Petal.Length, data=iris)
iris_lm
mean(iris$Petal.Length)
sd(iris$Petal.Length)
Petal.Length <- rnorm(150, mean=3.758,sd=1.765298)
irispre <- predict(iris_lm,data.frame(Petal.Length),interval="confidence")
irisy <- iris$Sepal.Length
irisbind <- as.data.frame(cbind(irispre,irisy))
ggplot(irisbind, aes(irisy, fit))+geom_point()
```



Iris 데이터를 활용하여 Sepal.Length를 종속변수로 하고 Petal.Length를 독립변수로 하는 iris_lm 생성

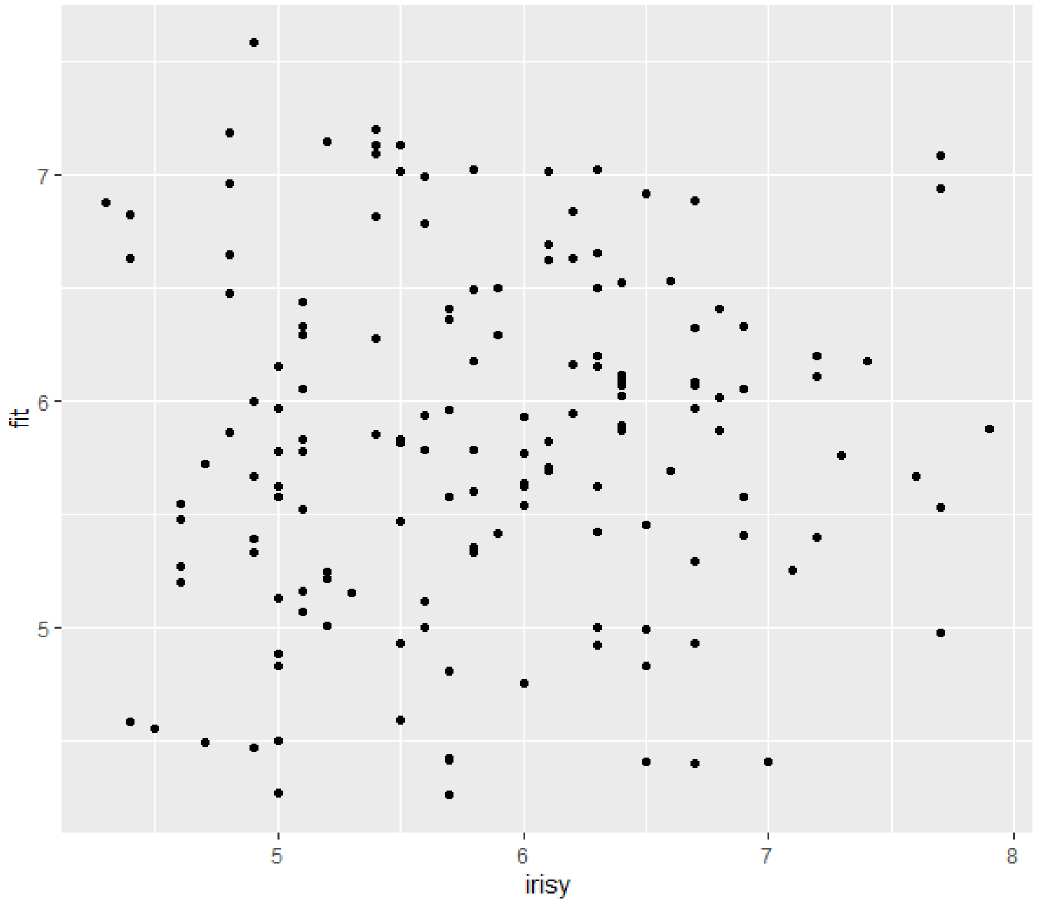
Petal.Length의 평균과 표준편차를 고려한 150개의 자료 구함

> Y값(관측된 종속변수 값)을 x축으로, predict값(새로운 독립변수에 대한 구간추정의 적합값)을 y축으로 한 그래프를 그려보았을 때 점들이 y=x 그래프를 형성하지 않음을 알 수 있음.

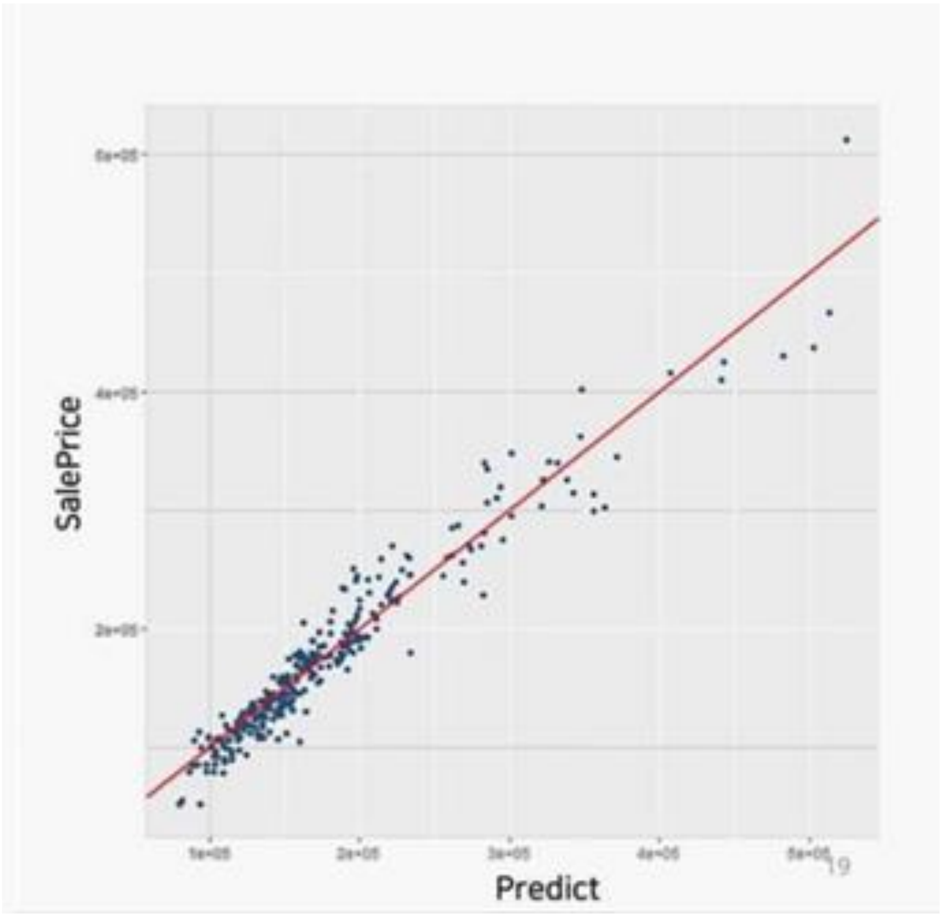
> 즉, 모델이 적합하지 않음을 확인할 수 있다.

+ y vs predict

< 모델이 적합하지 않은 경우 >



< 모델이 적합한 경우 >



+ Cp, AIC, BIC

SSR(Sum of Squared Residuals)를 활용한 모델의 적합성 판단

1. Cp 2. AIC 3. BIC

1. Cp

$$Cp = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} MSE(yihat)$$

: 예측력이 좋은 모형을 찾고 싶을 때 판단 기준이 됨.
(어떤 변수를 사용해야 예측력이 좋은 모형이 되는지 확인하는 기준)

n : 전체 데이터 개수
d : 변수의 개수
$\hat{\sigma}^2$: 예측된 분산

> Cp가 작을수록 예측력이 좋은 모델!

+ Cp, AIC, BIC

SSR(Sum of Squared Residuals)를 활용한 모델의 적합성 판단

1. Cp 2. AIC 3. BIC

2. AIC

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2) = Cp\frac{1}{\hat{\sigma}^2}$$

: Cp를 스케일링한 척도

n : 전체 데이터 개수
d : 변수의 개수
$\hat{\sigma}^2$: 예측된 분산

> AIC 가 작을수록 예측력이 좋은 모델!

+ Cp, AIC, BIC

SSR(Sum of Squared Residuals)를 활용한 모델의 적합성 판단

1. Cp 2. AIC 3. BIC

3. BIC

$$BIC = \frac{1}{n\hat{\sigma}^2}(RSS + log(n)d\hat{\sigma}^2) \qquad \mathrm{BIC} = \ln(n)k - 2\ln(\widehat{L}).$$

: 예측력이 좋은 모형을 찾고 싶을 때 판단 기준이 됨.
(Cp, AIC와 형태가 유사함)

N : 전체 데이터 개수
d, k : 변수의 개수
$\hat{\sigma}^2$: 예측된 분산
$\hat{L}$: 모델의 최대우도값

$$\hat{L} = p(x \mid \hat{\theta}) = \binom{x}{\hat{\theta}} p^{\hat{\theta}} (1-p)^{x-\hat{\theta}}$$ **($\hat{\theta}$ = 미지의 모수)**

n>7 인 경우에 log(n)이 2보다 커지기 때문에 BIC가 Cp보다 큰 값을 갖는 경우가 대다수
> Cp보다 모델의 선택 폭이 훨씬 적음. (정교함)

> BIC 가 작을수록 예측력이 좋은 모델!

비타민 '19.11.27.(수) Session

회귀진단 (Regression Diagnostics)

회귀모델의 적합성 판정+구간추정+다중공선성

2 조

이지선

우현우

그래도, 조금은.
그와 함께 이 팀의 무게를 같이 나누고 싶어.

–히노마루 스모

4. 독립변수 선정에서의 유의사항-"다중공선성"

다중공선성Multicollinearity이 발견된다면 이를 제거해주어야 한다.

다중공선성Multicollinearity

:
-독립변수들간의 **강한 상관관계**를 가지는 것. (완전한 선형관계인 경우: 완전공선성)
-독립 변수의 일부가 **다른 독립 변수의 조합으로 표현**될 수 있는 경우

이 사람은 농구를 잘 할 거야 <- 달리기 속도 + 키 + 팔길이 + 손크기 + 발크기 + 손가락길이 + 발가락길이

손 크면 손가락도 길던데.. / 발 크면 발가락도 길던데.. **(다중공선성)**→ *굳이 "손가락길이"와 "발가락길이"를 할 필요가 있을까?*

이 사람은 농구를 잘 할 거야 <- 달리기 속도 + 키 + 팔길이 + 손크기 + 발크기

문제가 되는 이유

:
-회귀계수의 최소제곱추정량(OLS를 통해 구한 추정값 $\hat{\beta}$)이 **합리적인 추정치를 제공해주지 못함**

-추정회귀계수($\hat{\beta}$)의 **분산이 매우 커지게 됨**
(회귀**계수값**이 존재하더라도 그 값이 **불안정**하여 **신뢰성있는 추정치를 얻을수 없음**)

4. 독립변수 선정에서의 유의사항-"다중공선성"

다중공선성Multicollinearity이 발견된다면 이를 제거해주어야 한다.

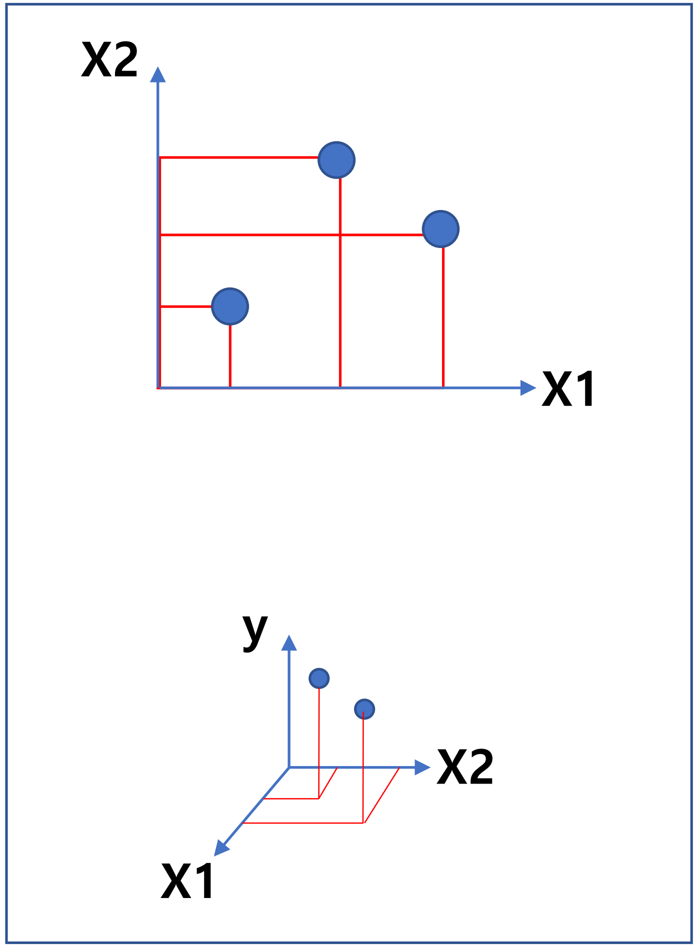
문제가 되는 이유

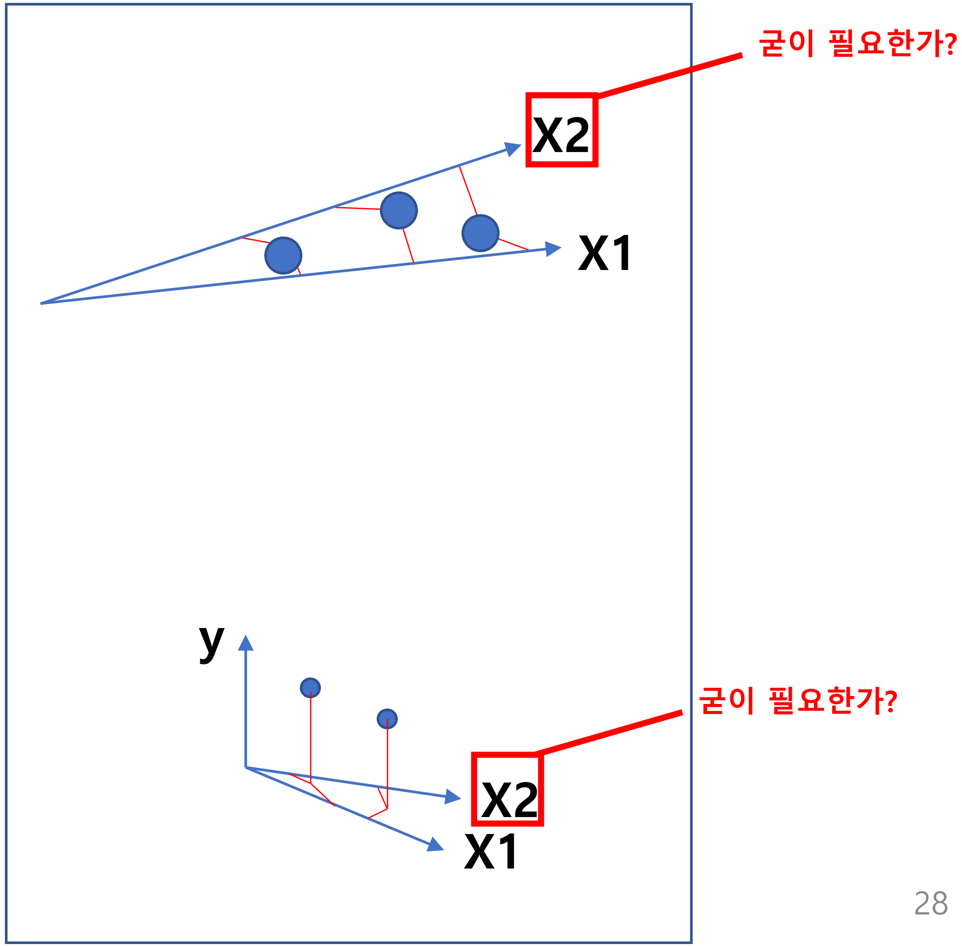
다중공선성 없는 경우(GOOD)

다중공선성 있는 경우(BAD)

2차원 관점:

3차원 관점:





4. 독립변수 선정에서의 유의사항-"다중공선성"

R 실습을 통해 회귀분석에서의 다중공선성Multicollinearity 문제를 해결하는 방식을 습득하자.

R 실습: 정수기 수리기사를 몇 명이나 고용할까?

엑셀 파일을 열어서 나온 Table을 복사하여 R의 데이터프레임으로 불러오자.


```
> #정수기 데이터프레임
> # [독립변수 x] purifier: 전월기준 정수기 총 대여수
> # [독립변수 x] old_purifier: 전월기준 10년 이상 노후 정수기 총 대여 수
> # [종속변수y] as_time: 당월 AS에 소요된 시간
> summary(purifier_df)
    purifier        old_purifier       as_time        new_purifier
 Min.   :168750   Min.   :26145   Min.   :19659   Min.   :128587
 1st Qu.:179708   1st Qu.:31805   1st Qu.:21200   1st Qu.:149519
 Median :194145   Median :36888   Median :22994   Median :157109
 Mean   :192377   Mean   :37215   Mean   :22896   Mean   :155162
 3rd Qu.:204000   3rd Qu.:43058   3rd Qu.:24317   3rd Qu.:162980
 Max.   :216375   Max.   :51552   Max.   :27070   Max.   :171885
```

현 상황

정수기 회사는 많은 정수기들을 사람들에게 대여해주고 있다.
해당 회사는 **정수기 수리 기사**를 고용하고 있는데, 현재 **어느 정도를 고용**하는게 좋을지 고민하고 있다.

따라서 **정수기 AS에 소요되는 시간이 얼마나 걸리는지** 예측하려 한다. (회귀분석)
왜냐하면 회사는 **1명의 수리기사가 정수기를 AS하는데 소요되는 시간을 알고 있기 때문**이다.

회귀분석을 진행함에 있어서, **정수기의 노후화 상태가 AS소요시간에 영향을 줄 것**이라 판단하여 이를 검토해본다.

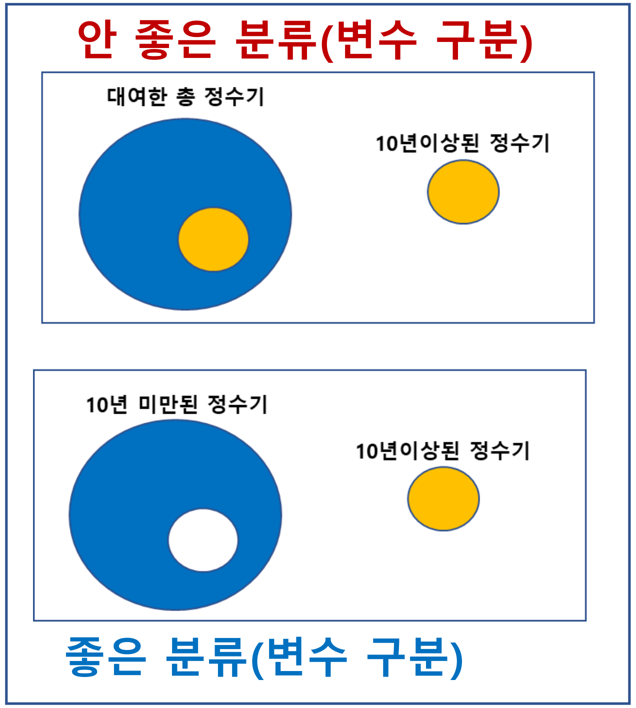
4. 독립변수 선정에서의 유의사항-"다중공선성"

R 실습을 통해 회귀분석에서의 다중공선성Multicollinearity 문제를 해결하는 방식을 습득하자.

R 실습: 정수기 수리기사를 몇 명이나 고용할까?

그런데, "전월 정수기 총 대여수"에는 이미 "10년 이상 정수기"가 포함되어 있다.
(따라서 Correlation=0.6042548)

*이러한 문제가 있음에도 불구하고, 높은 R^2 값들과 유의한 p-value. (**전형적인 다중공선성 증상 확인**)




```
> # 회귀분석에 있어서 독립변수간 상관성/포함성에 유의 (존재시 제거하여 모델의 왜곡 최소화)
> # 전월 정수기 총 대여 수 vs 10년 이상 정수기 상관성 분석
> cor(purifier_df$purifier, purifier_df$old_purifier)
[1] 0.6042548
> summary(lm(as_time~purifier+old_purifier,data=purifier_df))

Call:
lm(formula = as_time ~ purifier + old_purifier, data = purifier_df)

Residuals:
     Min       1Q   Median       3Q      Max 
-75.122 -14.427   -2.473   10.416  178.170 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.937e+02  1.077e+02   1.799   0.0828 .  
purifier     8.881e-02  6.742e-04 131.713   <2e-16 ***
old_purifier 1.510e-01  1.317e-03 114.609   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 40.97 on 28 degrees of freedom
Multiple R-squared:  0.9996,	Adjusted R-squared:  0.9996 
F-statistic: 3.837e+04 on 2 and 28 DF,  p-value: < 2.2e-16
```

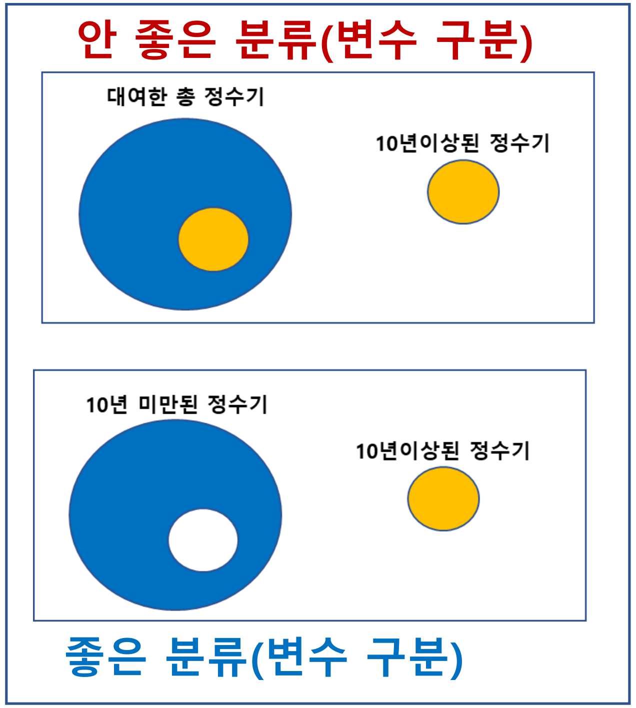
4. 독립변수 선정에서의 유의사항-"다중공선성"

R 실습을 통해 회귀분석에서의 다중공선성Multicollinearity 문제를 해결하는 방식을 습득하자.

R 실습: 정수기 수리기사를 몇 명이나 고용할까?

대여한 정수기를 교집합부분이 없도록 나누자. (10년 이상 / 10년 이상)
-> correlation = 0.1151678 확인! Good (다중공선성 해결)

이후 다시 회귀분석을 시행.




```
> # 10년 미만 정수기 vs 10년 이상 정수기 상관성 분석
> cor((purifier_df$purifier-purifier_df$old_purifier), purifier_df$old_purifier)
[1] 0.1151678
```

```
> #변수 재정의
> # [독립변수 x] 전월기준 10년 이상 노후 정수기 총 대여 수
> # [독립변수 x] 전월기준 10년 미만 노후 정수기 총 대여 수
> # [종속변수y] as_time: 당월 AS에 소요된 시간
>
> # '전월기준 10년 미만 노후 정수기 총 대여수' 변수(new_purifier) 추가
> str(purifier_df)
'data.frame':	31 obs. of  4 variables:
 $ purifier    : num  168750 171450 172800 174000 174810 ...
 $ old_purifier: num  33750 42863 31104 40020 26222 ...
 $ as_time     : num  20453 21850 20214 21660 19659 ...
 $ new_purifier: num  135000 128587 141696 133980 148588 ...
> purifier_df$new_purifier <- purifier_df$purifier-purifier_df$old_purifier
> str(purifier_df)
'data.frame':	31 obs. of  4 variables:
 $ purifier    : num  168750 171450 172800 174000 174810 ...
 $ old_purifier: num  33750 42863 31104 40020 26222 ...
 $ as_time     : num  20453 21850 20214 21660 19659 ...
 $ new_purifier: num  135000 128587 141696 133980 148588 ...
```

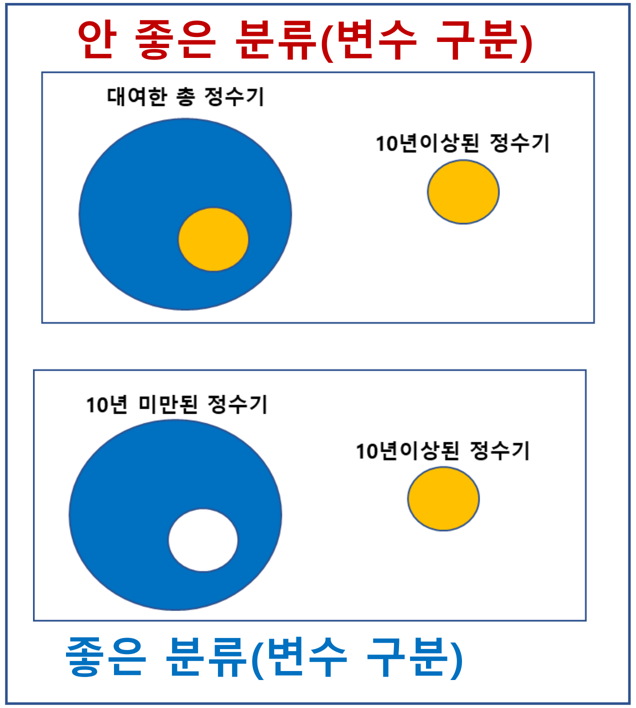
4. 독립변수 선정에서의 유의사항-"다중공선성"

R 실습을 통해 회귀분석에서의 다중공선성Multicollinearity 문제를 해결하는 방식을 습득하자.

R 실습: 정수기 수리기사를 몇 명이나 고용할까?

대여한 정수기를 교집합부분이 없도록 나누자. (10년 이상 / 10년 이상)
-> correlation = 0.1151678 확인! Good (다중공선성 해결)

이후 다시 회귀분석을 시행.




```
> #회귀분석 수행
> # [독립변수 x] 10년 미만 정수기(new_purifier), 10년 이상 정수기(old_purifier)
> # [종속변수y] as_time: 당월 AS에 소요된 시간
> lm_result<-lm(as_time~new_purifier+old_purifier,data=purifier_df)
> summary(lm_result)

Call:
lm(formula = as_time ~ new_purifier + old_purifier, data = purifier_df)

Residuals:
    Min      1Q  Median      3Q     Max 
-75.122 -14.427  -2.473  10.416 178.170 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.937e+02  1.077e+02   1.799   0.0828 .  
new_purifier 8.881e-02  6.742e-04 131.713   <2e-16 ***
old_purifier 2.398e-01  1.057e-03 226.933   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 40.97 on 28 degrees of freedom
Multiple R-squared:  0.9996,	Adjusted R-squared:  0.9996 
F-statistic: 3.837e+04 on 2 and 28 DF,  p-value: < 2.2e-16

> # p-value<0.05이므로 유의한 모델, R^2=0.9996으로 높은 설명력을 가지는 회귀모델
> # y=193.7+0.08881(x1)+0.2398(x2)
> # 10년미만정수기(x1) 1대당 0.0881시간 소요, 10년이상정수기(x2) 1대당 0.2398시간 소요
```

4. 독립변수 선정에서의 유의사항-"다중공선성"

R 실습을 통해 회귀분석에서의 다중공선성Multicollinearity 문제를 해결하는 방식을 습득하자.

R 실습: 정수기 수리기사를 몇 명이나 고용할까?

10년 미만 정수기 수와 10년 이상 정수기 수를 회사에서 알고 있다. (x값들을 알고 있다)
이 값들을 우리가 만든 회귀식에 대입해보자.
이를 통해, "AS소요시간"(y값)을 예측할 수 있다.

```
> #10년미만 정수기가 300,000대, 10년이상 노후 정수기가 70,000대
> #회귀분석 값에 대입할 독립변수값 설정(데이터 프레임)
> input_predict<-data.frame(new_purifier=300000,old_purifier=70000)
> #회귀모델에 독립변수값 저장후 출력: AS시간이 43619시간 소요
> predict_as_time<-predict(lm_result,input_predict)
> predict_as_time
       1
43619.54
```

AS기사가 한달간 소요하는 AS시간을 기반으로,
1개월동안 필요한 AS기사 수를 구하자.

```
> #AS기사 1명이 한달간 처리하는 AS시간=8시간*20일=160시간
> #해당수치로 총AS시간을 나눠주면, 272.6221(약273) 이라는 값이 나옴.(필요한 AS기사 수)
> predict_as_time/(8*20)
       1
272.6221
```

4. 독립변수 선정에서의 유의사항-"다중공선성"

R 실습을 통해 회귀분석에서의 다중공선성Multicollinearity 문제를 해결하는 방식을 습득하자.

R 실습: 정수기 수리기사를 몇 명이나 고용할까?

구간 추정(신뢰구간 찾기)을 해보자.
(모평균의 신뢰성을 가늠)

신뢰수준 95%[유의수준α=0.05, 신뢰수준=100(1- α)]에서 예상되는 AS소요시간의 범위를 살펴보자. (= 신뢰구간을 구하자)

```
> #구간 추정(43619시간을 기준으로 95%의 유의수준: 43414시간~43824시간)
> predict_as_time <-predict(lm_result, input_predict, interval="confidence",level=0.95)
> predict_as_time
       fit      lwr     upr
1 43619.54 43414.58 43824.5
```

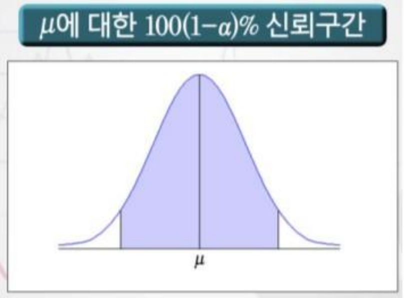
→ <u>AS소요시간</u>은 **약 43414.58시간 ~ 43824.5시간** 쯤 되는구나!

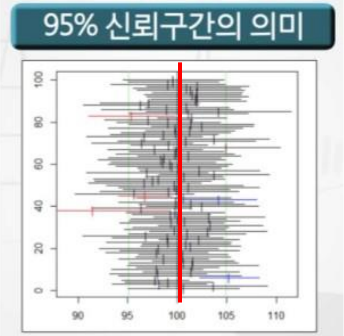
→ <u>해석:</u>

동일한 방법으로 100번 표본추출했을 때에 나온 신뢰구간 중..
모평균(43619.54시간)을 포함한 신뢰구간은 95개정도군!

*("<u>모평균을 포함할 확률이 95%가 되는 구간</u>"은 **틀린 해석**)*

***모평균**이 뚫고가는 가로선(신뢰구간)이 <u>95개!</u>*
<u>(5개는 빨간 직선과 만나지 않아)</u>





4. 독립변수 선정에서의 유의사항-"다중공선성"

다중공선성Multicollinearity 문제를 확인하는 방안에 대해 알아보자.

다중공선성 확인방법1.1: 상관계수correlation coefficient

변수간의 **상관계수correlation coefficient**를 살펴본다 (0.5 이상 또는 0.7 이상을 다중공선성 기준으로 대체로 사용)

"10년이상 된 정수기 수"와 "총 정수기 수"의 상관계수correlation coefficient **확인**

```
> # 전월 정수기 총 대여 수 vs 10년 이상 정수기 상관성 분석
> cor(purifier_df$purifier, purifier_df$old_purifier)
[1] 0.6042548
```

[표 11] 분석지표

상관관계 계수	해　석
0.0 ~ 0.2	상관관계가 거의 없다
0.2 ~ 0.4	상관관계가 다소 있다
0.4 ~ 0.6	상관관계가 다소 높다
0.6 ~ 0.8	상관관계가 높다
0.8 ~ 1.0	상관관계가 아주 높다

4. 독립변수 선정에서의 유의사항-"다중공선성"

다중공선성Multicollinearity 문제를 확인하는 방안에 대해 알아보자.

다중공선성 확인방법1.2: 산점도Scatter Plot

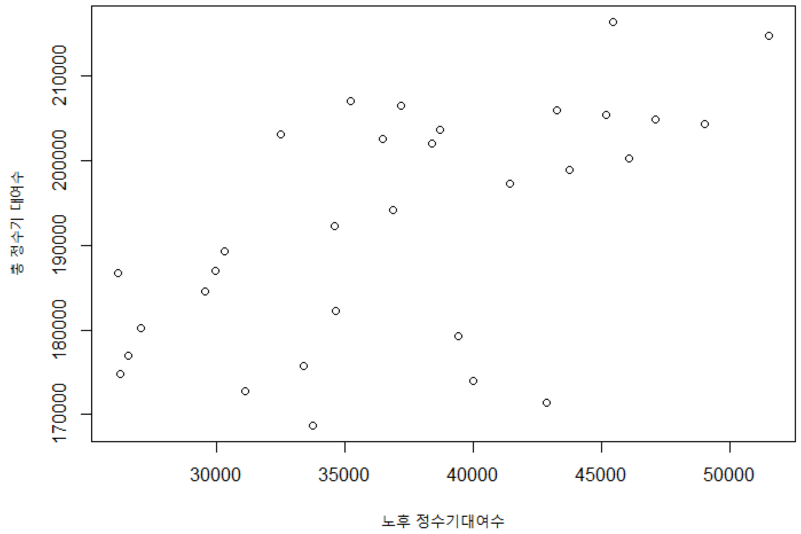
산점도를 통해서도 살펴보자.

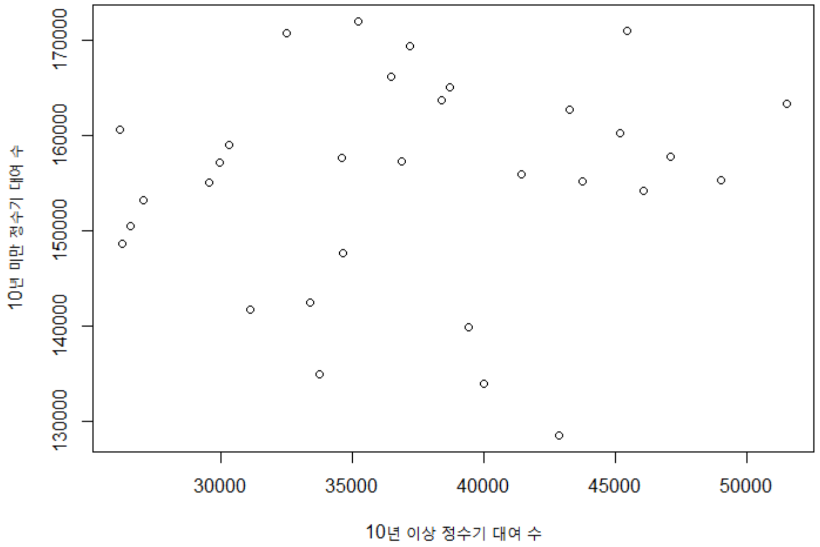
"10년이상 된 정수기 수"와 "총 정수기 수"의 산점도 보기 ~"우상향의 형태"가 보인다

```
> plot(purifier_df$old_purifier, purifier_df$purifier, xlab="노후 정수기대여수", ylab="총 정수기 대여수")
```

"10년이상 된 정수기 수"와 "10년미만 된 정수기 수"의 산점도 보기 ~"규칙성"이 안 보인다

```
> plot(purifier_df$old_purifier, purifier_df$new_purifier, xlab="10년 이상 정수기 대여 수", ylab="10년 미만 정수기 대여 수")
```





4. 독립변수 선정에서의 유의사항-"다중공선성"

다중공선성Multicollinearity 문제를 확인하는 방안에 대해 알아보자.

다중공선성 확인방법2: 분산팽창지수(Variation Index Factor, VIF)

분산팽창지수=분산팽창인자(VIF):

$$VIF_i = \frac{1}{1-R_i^2}$$

Xi가 다른 변수들(X)에 의해 설명이 잘 된다
→ R_i^2 가 크다 → VIF가 크다

→일반적으로 이 값이 **10 이상**이면
그 독립변수(X_i)는 **다중공선성** 문제를 발생!
(기준을 3이나 5로 잡는 경우도 있음)

```
library(car)
```

"car"패키지에 내장!

```
## Loading required package: carData
```

```
vif(fit1)
```

R코드: vif(데이터명)

```
##  Sepal.Width Petal.Length  Petal.Width
##     1.270815    15.097572    14.234335
```

```
# Petal.Length와 Petal.Width에서 높은 다중공선성 발견
```

Appendix

다중회귀분석 $Y \leftarrow X_1, \cdots, X_p$ 을 할 때 i 번째 독립변수에 대한 다중회귀분석 $X_i \leftarrow X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_p$ 의 중회귀계수를 R_i^2 라고 두자. $\text{VIF}_i := \frac{1}{1-R_i^2}$ 를 **분산팽창인자**Variance Inflation Factor라고 한다.

쉬운 설명: y=b0+b1x1+…+bpXp 에 대해서
X_i에 대해 다중공선성을 확인하자.

이때, X_i를 종속변수로 하고, 나머지 X값들을 독립변수로 하는 회귀식을 세우자.
이때 이 새로운 회귀식에 대한 결정계수를 R_i^2라 한다.

*참고(상태지수, 분산팽창인자) 링크들
https://m.blog.naver.com/PostView.nhn?blogId=kwonkise&logNo=220619258389&proxyReferer=https%3A%2F%2Fwww.google.com%2F
http://elearning.kocw.net/contents4/document/lec/2013/Konkuk/Choijaeheon/11.pdf
https://maengdev.tistory.com/106
https://byzun0.tistory.com/90
https://ai-times.tistory.com/268
http://wolfpack.hnu.ac.kr/2015_Spring/LM/LM_%EB%8B%A4%EC%A4%91%EA%B3%B5%EC%84%A0%EC%84%B1.pdf

4. 독립변수 선정에서의 유의사항-"다중공선성"

다중공선성Multicollinearity 문제를 확인하는 방안에 대해 알아보자.

다중공선성 확인방법3: 상태지수Condition Index

상태지수 Condition Index:

$$C_i = \sqrt{\frac{\lambda_{max}}{\lambda_i}}$$

상태지수 C_i는..

(option; 고유값 0.01 이하일 때)
상태지수 10 이상 → 공선성 발생 가능성 있음

(option; 고유값 0.001 이하일 때)
상태지수 100 이상 → 공선성이 강하게 발생 가능성 있음

**30을 기준으로 하기도 함*

Appendix

**[모델에서 가장 변동을 설명 잘하는 요인(주성분)의 설명력]을 [다른 요인들이 변동을 설명하는 정도(설명력)]로 나눈다*

**그런데.. 이 나눈 값(비율)이 굉장히 크네?*
=주성분의 설명력이 굉장히 크네?
=상태지수가 크네?
→ 주성분은 공선성을 가지고 있군!

- λ =고유값=고유치=eigen value

*"고유값"은..
-요인의 설명력을 의미
-요인이 얼마나 변수들의 분산을 잘 설명하는가

차이점 check!

분산팽창인자VIF
: 각각의 모든 독립변수의 공산성 확인
vs
상태지수 Condition Index
: 가장 설명력 높은 독립변수의 공산성 확인

4. 독립변수 선정에서의 유의사항-"다중공선성"

다중공선성Multicollinearity 문제를 해결하는 방안에 대해 알아보자.

다중공선성 해결방법

1. **변수선택법 Variable Selection**
2. **PCA(Principal Component Analysis, 주성분분석)**
3. 표본의 수(Sample Size)를 늘려보기
4. 변수를 변형시키거나, 새로운 관측치(New Sample)를 이용해보기

4. 독립변수 선정에서의 유의사항-"다중공선성"

다중공선성Multicollinearity 문제를 해결하는 방안에 대해 알아보자.

다중공선성 해결방법 (간략 소개) **Appendix**

1. 변수선택법 Variable Selection

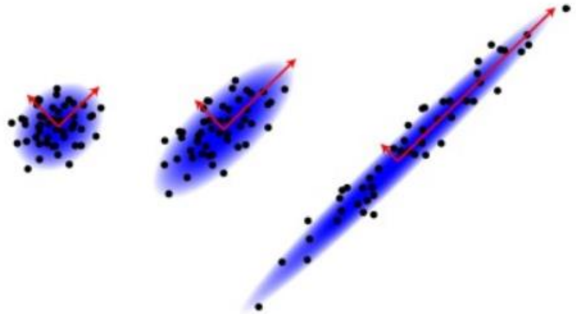
목적:
변수가 많을수록 설명력은 커진다(오류가 줄어든다). 그러나 모형이 복잡해지고 쓸모없는 변수가 들어갈 수 있다.
따라서 **적절한 변수의 수를 결정**하기 위해 **유효한 변수만을 채택**하는 방법을 '변수 선택법'이라 함.

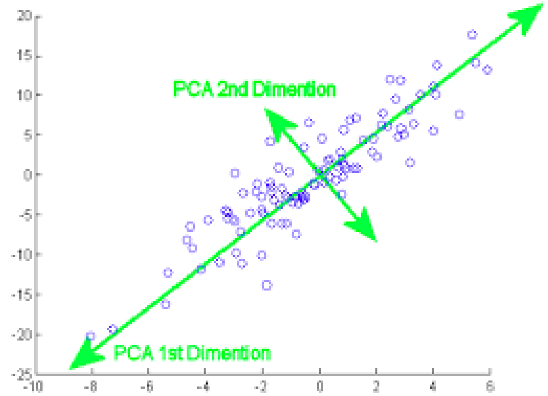
-Forward Selection(전진선택법)
모델에 모든 변수를 다 없앤 상태에서
하나씩 변수를 넣는 방식
(그래서 그 변수를 넣으면 모델 설명력이 좋아지는가로 결정)

-Backward Selection(후진제거법)
모델에 모든 변수를 다 넣은 상태에서
하나씩 변수를 빼는 방식
(그래서 변수를 빼면 모델 설명력이 좋아지는가로 결정)

-Stepwise Regression(단계적회귀법)
위 2가지 방식의 혼합.
모델에 변수를 하나씩 넣어보고, 빼보고를 반복하는 방식

2. PCA(Principal Component Analysis, 주성분분석)





#개념:

-상관관계가 있는 변수들을 선형 결합하여 변수를 축약하는 기법.

*PCA는 데이터 하나 하나에 대한 성분을 분석하는 것이 아니라,
여러 데이터들이 모여 하나의 분포를 이룰 때 이 분포의 주성분을 분석해주는 방법

PCA (Principal Components Analysis, 주성분분석)	주성분 = f (변수)
FA (Factor Analysis, 요인분석, 인자분석)	변수 = f (요인)

https://niceguy1575.tistory.com/59

4. 독립변수 선정에서의 유의사항-"다중공선성"

다중공선성Multicollinearity 문제 Process

요약정리

다중공선성Multicollinearity

: 모델에서 "독립변수"간의 높은 상관 관계(연관성, 관련성)

→ 모델에 합리적 추정/분산측면에서 방해된다(신뢰성, 불안정성) + 불필요

(ex: 독립변수-종속변수가 양의 상관관계인데, 모델에서 독립변수는 음의 계수를 가짐)
(ex: R-squared가 이상하게 너무 높은 경우)

진단방법

정성적
- 산점도 ~*모델의 계수와 산점도의 상관성의 부호가 다른것으로도 추정 가능*

정량적
- 상관계수 : (0.5 또는 0.7 이상이면 check!) ~*모델의 계수와 상관계수의 부호가 다른것으로도 추정 가능*
- 분산팽창지수(VIF) : 각각의 모든 독립변수의 공산성 확인 (10이상이면 check!)
- 상태지수(Condition Index) : 가장 설명력 높은 독립변수의 공산성 확인 (10이상 또는 30이상 또는 100이상 이면 check!)

↓

해결방법
- 변수선택법 : forward / backward / stepwise
- PCA : 독립변수가 비슷한 애들끼리 묶기
- 표본 수(Sample Size) 늘려보기
- 변수변형 or 새로운 표본(New Sample)로 다시 해보기

비타민 19.11.27.(수) Session

회귀진단 (Regression Diagnostics)

회귀모델의 적합성 판정+구간추정+다중공선성

-End of Document-

2 조

이지선

우현우