R을 이용한 Conditional Process Analysis

I. 회귀에서 조절까지

문건웅

2019/6/1

회귀분석

- 관찰된 연속형 변수들에 대해 두 변수 사이의 모형을 구한뒤 적합도를 측정해 내는 분석 방법
- 단순회귀분석(simple regression analysis) :
  - 하나의 종속변수와 하나의 독립변수 사이의 관계를 분석
- 다중회귀분석(multiple regression) :
  - 하나의 종속변수와 여러 독립변수 사이의 관계를 분석
- 상호작용이 있는 회귀분석 :
  - 독립변수들 사이에 상호작용이 있는 경우
  - 하나의 변수(조절변수)의 값이 변화함에 따라 다른 독립변수와 종속변수 사이의 회귀선의 기울기가 변하는데 이것을 조절효과라고 할 수 있다.

자동차의 연비

COOPER Mini

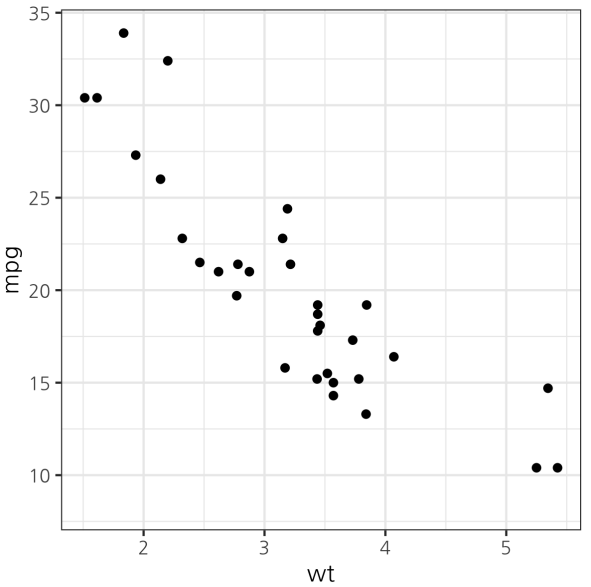
- 공차중량 1,230kg
- 배기량 1,496cc
- 최대출력 116마력
- 연비 16.5km/L

BENZ S600 MAYBACH

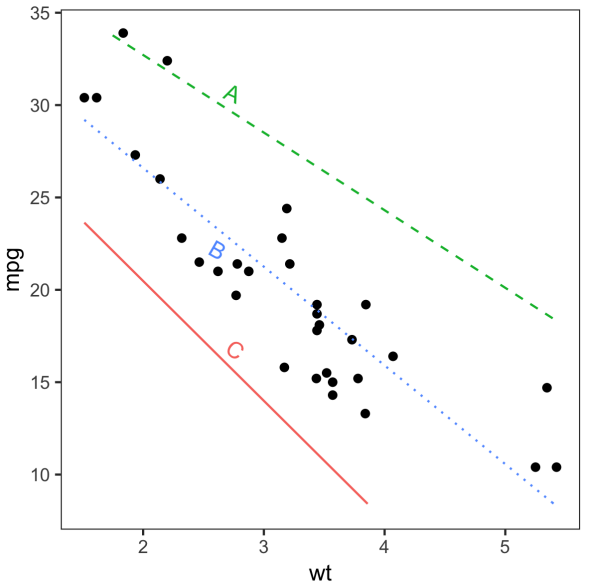
- 공차중량 2,345kg
- 배기량 3,982cc
- 최대출력 630마력
- 연비 8.1km/L

자동차의 연비와 공차중량의 관계

- 데이터 mtcars (1974, Motor Trend)



두 변수의 관계를 직선으로 나타내보면?



단순회귀분석

```
fit=lm(mpg ~ wt, data=mtcars)
summary(fit)
```

```
Call:
lm(formula = mpg ~ wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5432 -2.3647 -0.1252  1.4096  6.8727

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
wt           -5.3445     0.5591  -9.559 1.29e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.046 on 30 degrees of freedom
Multiple R-squared:  0.7528,    Adjusted R-squared:  0.7446
F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

회귀분석 결과 요약

회귀 분석의 결과 회귀선의 기울기는 -5.34이고 y절편은 37.29이다. 즉, 회귀직선을 식으로 나타내면 다음과 같다.

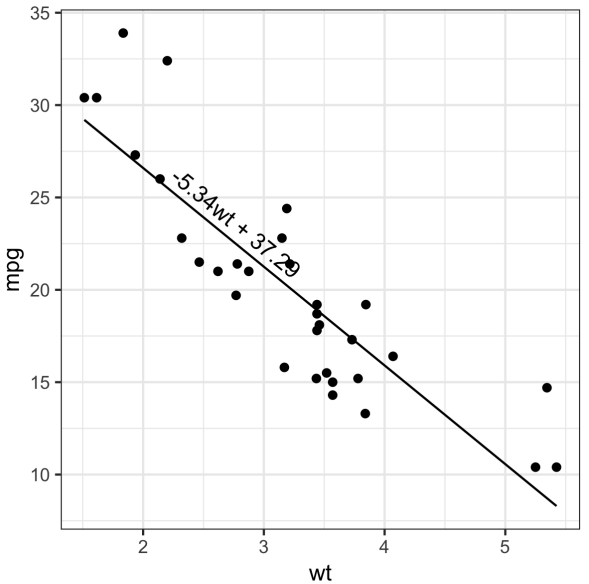
> $$mpg = -5.34wt + 37.29$$

위 공식을 일반화하여 종속변수를 Y, 독립변수를 X, 회귀선의 y절편을 a, 기울기를 b라고 하면 반응변수 Y의 추정치(yhat, $\hat{Y}$)는 다음과 같다.

> $$\hat{Y} = a + bX$$

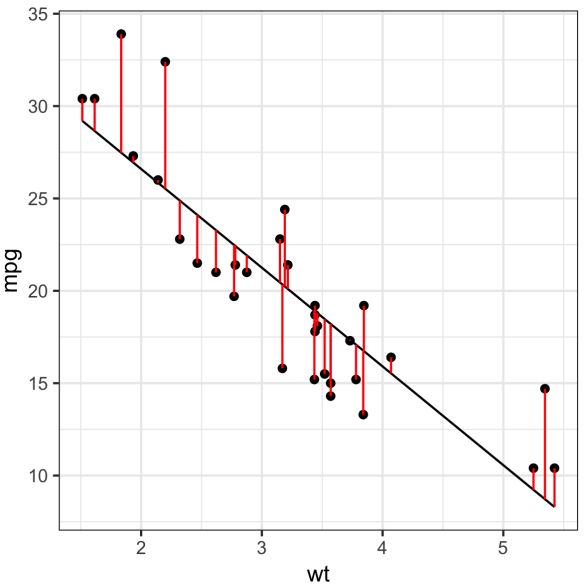
산점도와 회귀직선

```
require(predict3d)
ggPredict(fit)
```



회귀식과 오차

```
ggPredict(fit,show.text=FALSE, show.error=TRUE)
```



보통최소제곱회귀

반응변수 y의 i번째 관측치 y_i는 다음과 같이 나타낼 수 있다.

> $$y_i = a + bx_i + \varepsilon_i$$

여기서 $a + bx_i$는 i번째 반응변수의 추정치 $\hat{y}_i$이므로 i번째 잔차 ε_i 는 다음과 같다.

> $$\varepsilon_i = y_i - \hat{y}_i$$

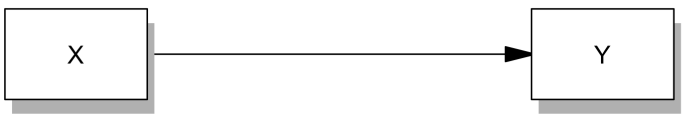
이들 잔차는 회귀선의 적합도를 나타내는데 잔차의 합은 0이 되므로 잔차의 제곱합을 사용하며 보통의 OLS(ordinary least square) 회귀에서는 잔차의 제곱합이 최소가 되도록 기울기와 y절편을 추정한다.

단순회귀분석의 개념적모형

R로 단순회귀분석을 할 경우 lm()함수의 formula로 Y ~ X 와 같이 사용하며 이를 개념적모형으로 나타내면 다음과 같다.

> R formula: Y ~ X

```
require(processR)
pmacroModel(0,radx=0.1,rady=0.07,ylim=c(0.1,0.6))
```

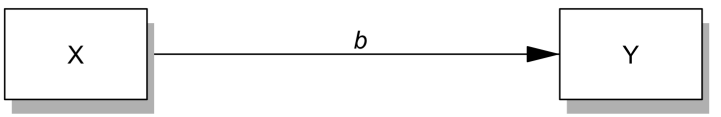


단순회귀분석의 통계적 모형

통계적모형으로 나타낼 때는 기울기를 같이 표시해준다. 이때 기울기를 b라고 하면 Y ~ b * X 로 표시할 수 있다. lavaan 패키지의 sem()함수를 이용하여 분석을 할 경우 이와 같이 기울기를 지정해주는 것이 좋다.

> Model syntax: Y ~ b * X

```
statisticalDiagram(0,radx=0.1,rady=0.07,,ylim=c(0.1,0.6))
```



다중회귀분석 - 상호작용이 없는 경우

- mtcars 데이터의 vs 변수는 engine이 “V-shape”인 경우 0, “straight”인 경우 1로 되어 있다.
- 연비(mpg)의 설명변수로 공차중량(wt)과 함께 vs를 설명변수로 하는 회귀모형을 만든다

```
mtcars$engine=factor(mtcars$vs,labels=c("V-shape","Straight"))
fit1=lm(mpg ~ wt + engine, data = mtcars)
summary(fit1)
```

```
Call:
lm(formula = mpg ~ wt + engine, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7071 -2.4415 -0.3129  1.4319  6.0156

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     33.0042     2.3554  14.012 1.92e-14 ***
wt              -4.4428     0.6134  -7.243 5.63e-08 ***
```

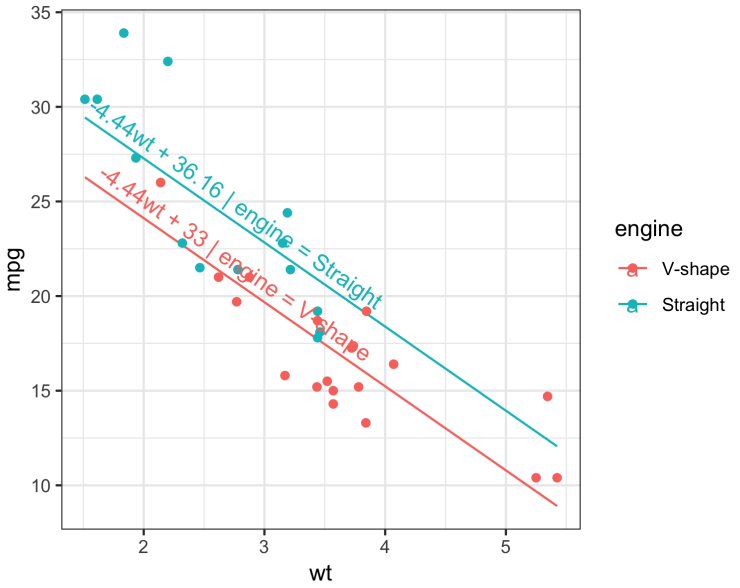
회귀모형 해석

mpg를 y축으로 wt를 x 축으로 하는 회귀선의 기울기는 -4.44로 동일하나 y절편은 engine 에 따라 달라진다.

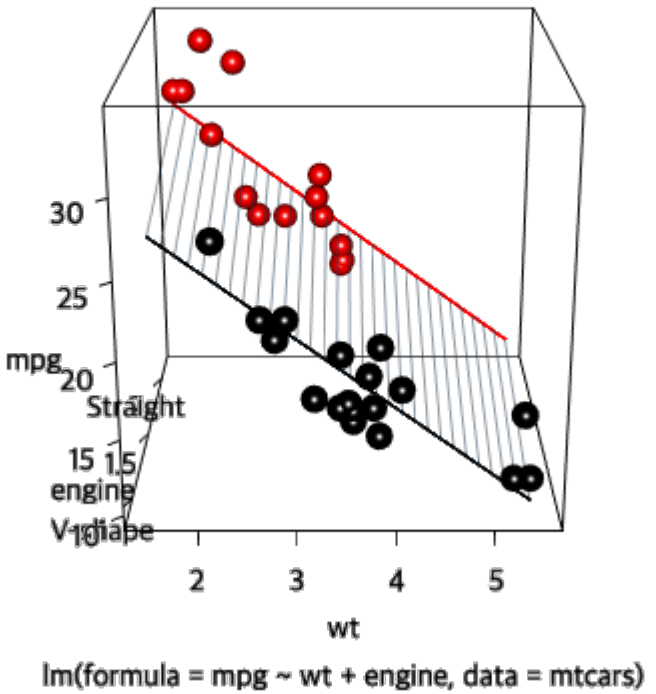
- engine이 V-shape 인 경우 $intercept \approx 33.0$
- engine이 Straight 인 경우 $intercept \approx 33.00042 + 3.1544 \approx 36.16$

이 모형을 시각화하면 다음과 같다.

```
ggPredict(fit1)
```



```
require(predict3d)
predict3d(fit1,radius=0.5)
```

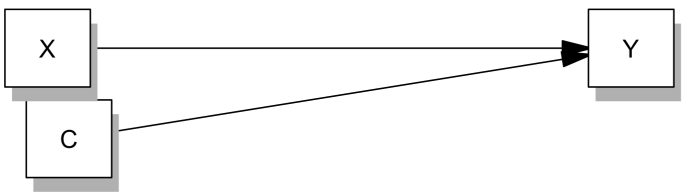


개념적모형

상호작용이 없는 다중회귀분석의 경우 다음과 같은 R formula를 사용한다. 이때 C는 공변량(covariate)를 뜻한다.

> R formula: Y ~ X + C

```
pmacroModel(0,covar=list(name="C",site=list("Y")),ylim=c(0.1,0.6))
```

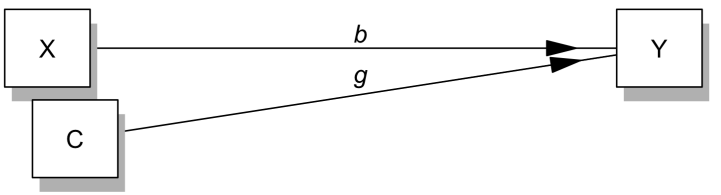


통계적모형

- 공변량의 회귀계수는 Y에 영향을 미치는 공변량의 경우 g1, g2,...로 붙인다.
- M에 영향을 미치는 공변량의 경우 f1, f2,...로 붙인다.

> Model syntax: Y ~ b * X + g * C

```
statisticalDiagram(0,radx=0.06,rady=0.06,covar=list(name="C",site=lis
```



상호작용이 있는 다중회귀모형(1)

- 상호작용이 있는 경우 X:W의 형식으로 표기
- Y가 반응변수이고 X, W가 설명변수이고 X와 W의 상호작용이 있는 경우의 모형을 R formula로 나타내면 다음과 같다.

> Y ~ X + W + X:W

위의 formula를 간단하게 Y ~ X*W 로도 쓸 수 있다. *는 모든 가능한 상호작용을 뜻한다.

> X * Z = X + Z + X:Z
>
> A * B * C = A + B + C + A:B + B:C + A:C + A:B:C

상호작용이 있는 다중회귀분석

```
fit2=lm(mpg ~ wt*engine,data=mtcars)
summary(fit2)
```

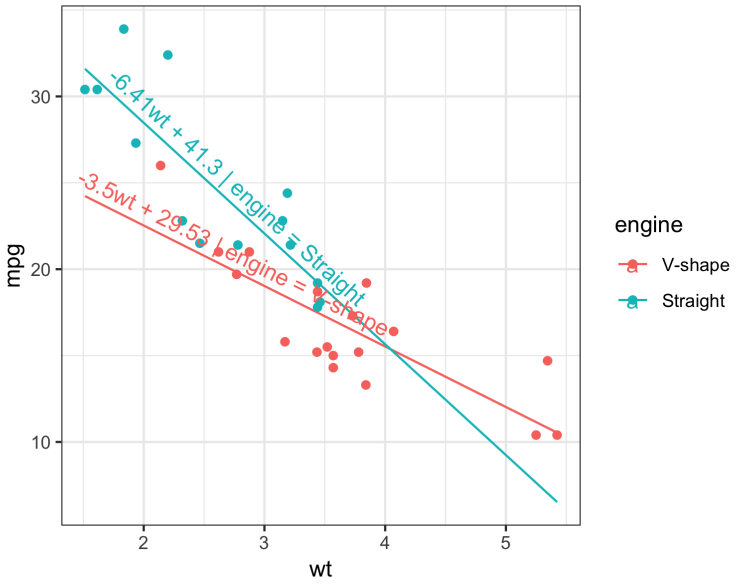
```
Call:
lm(formula = mpg ~ wt * engine, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.9950 -1.7881 -0.3423  1.2935  5.2061 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        29.5314     2.6221  11.263 6.55e-12 ***
wt                 -3.5013     0.6915  -5.063 2.33e-05 ***
engineStraight     11.7667     3.7638   3.126   0.0041 ** 
wt:engineStraight  -2.9097     1.2157  -2.393   0.0236 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.578 on 28 degrees of freedom
Multiple R-squared:  0.8348,    Adjusted R-squared:  0.8171 
F-statistic: 47.16 on 3 and 28 DF,  p-value: 4.497e-11
```

회귀모형의 기울기와 y절편

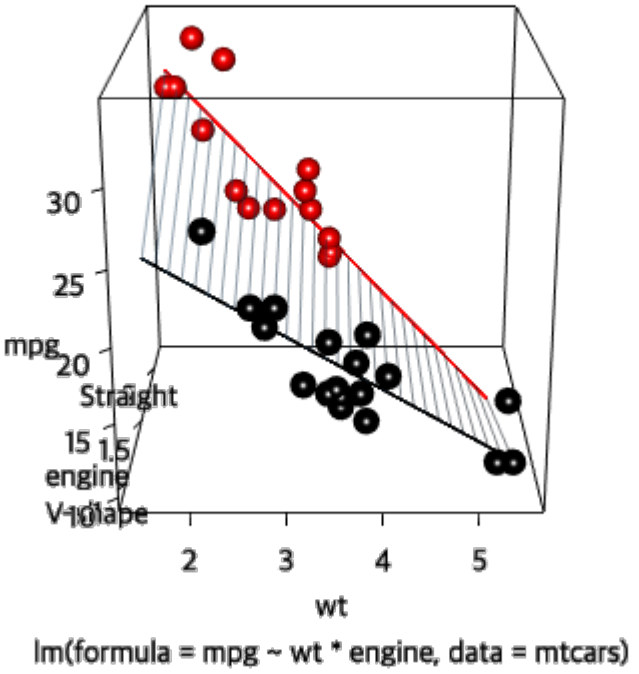
이 모형에서 wt와 engine의 상호작용 wt:engine은 통계적으로 의미있다(p = 0.0236). 이 모형에서 wt를 설명변수로 engine를 조절변수의 역할을 한다고 생각해 본다. mpg와 wt사이의 회귀식의 기울기와 y절편은 vs 의 값에 따라 달라진다.

- y 절편
  - engine이 V-shaped 인 경우 $intercept \approx 29.5314$
  - engine이 Straight 인 경우 $intercept \approx 29.5314 + 11.7667 \approx 41.3$
- 기울기
  - engine이 V-shaped 인 경우 $slope \approx -3.5013$
  - engine이 Straight 인 경우 $slope \approx -3.5013 - 2.9097 \approx 6.41$

```
ggPredict(fit2)
```



```
predict3d(fit2,radius=0.5)
```



상호작용이 있는 다중회귀모형(2)

- 조절변수가 연속형변수인 경우

```
fit3 = lm( mpg ~ wt*hp, data=mtcars)
summary(fit3)
```

```
Call:
lm(formula = mpg ~ wt * hp, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.0632 -1.6491 -0.7362  1.4211  4.5513 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 49.80842    3.60516  13.816 5.01e-14 ***
wt          -8.21662    1.26971  -6.471 5.20e-07 ***
hp          -0.12010    0.02470  -4.863 4.04e-05 ***
wt:hp        0.02785    0.00742   3.753 0.000811 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.153 on 28 degrees of freedom
```

기울기와 y절편 계산

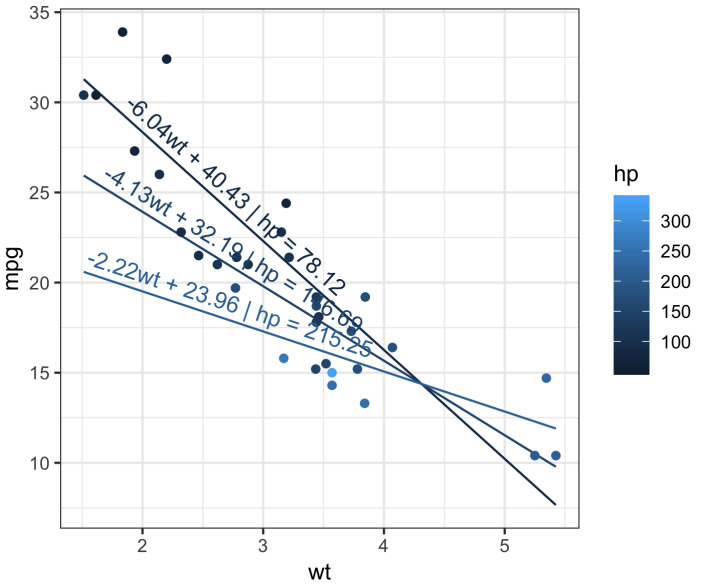
wt의 평균과 평균-표준편차, 평균+ 표준편차를 계산해보면 다음과 같다.

```
mean(mtcars$hp, na.rm=TRUE) + c(-1,0,1)* sd(mtcars$hp, na.rm=TRUE)
```

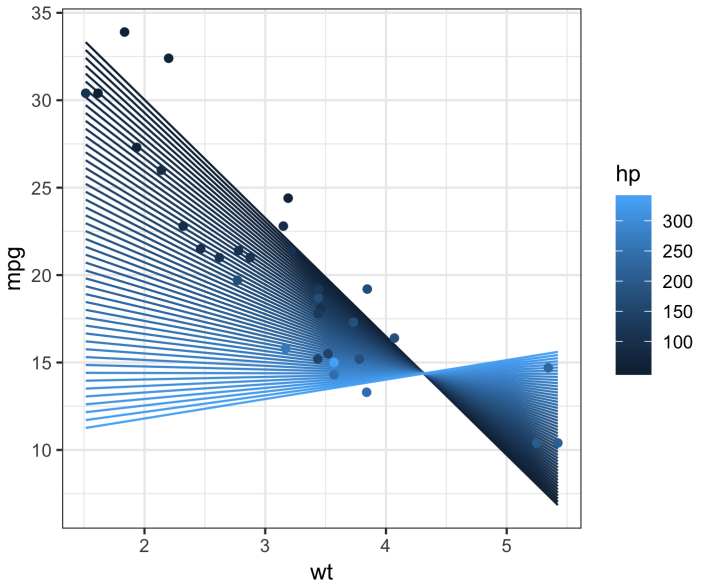
```
[1]  78.12463 146.68750 215.25037
```

- y 절편
  - hp가 78.12 인 경우 $intercept \approx 49.81 - 0.12 \times 78.12 \approx 40.43$
  - hp가 146.69 인 경우 $intercept \approx 49.81 - 0.12 \times 146.69 \approx 32.19$
  - hp가 215.25 인 경우 $intercept \approx 49.81 - 0.12 \times 215.25 \approx 23.96$
- 기울기
  - hp가 78.12 인 경우 $slope \approx -8.22 + 0.028 \times 78.12 \approx -6.04$
  - hp가 146.69 인 경우 $slope \approx -8.22 + 0.028 \times 146.69 \approx -4.13$
  - hp가 215.25 인 경우 $slope \approx -8.22 + 0.028 \times 215.25 \approx -2.22$

```
ggPredict(fit3)
```

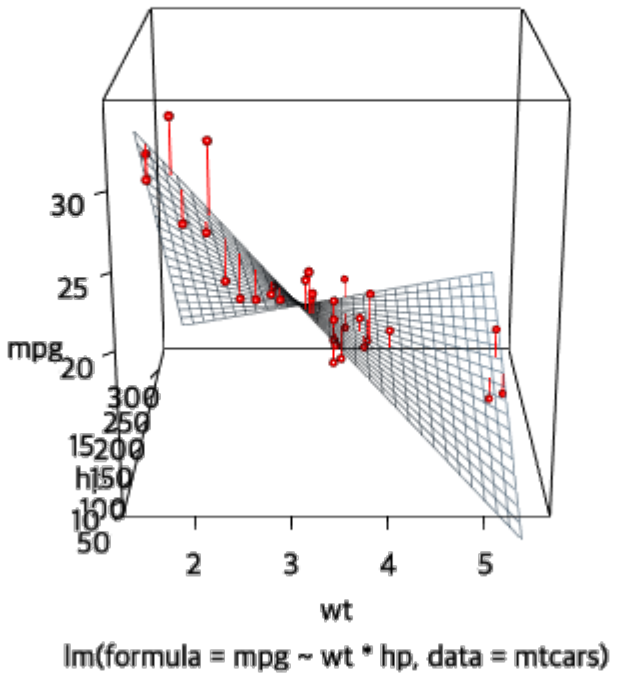


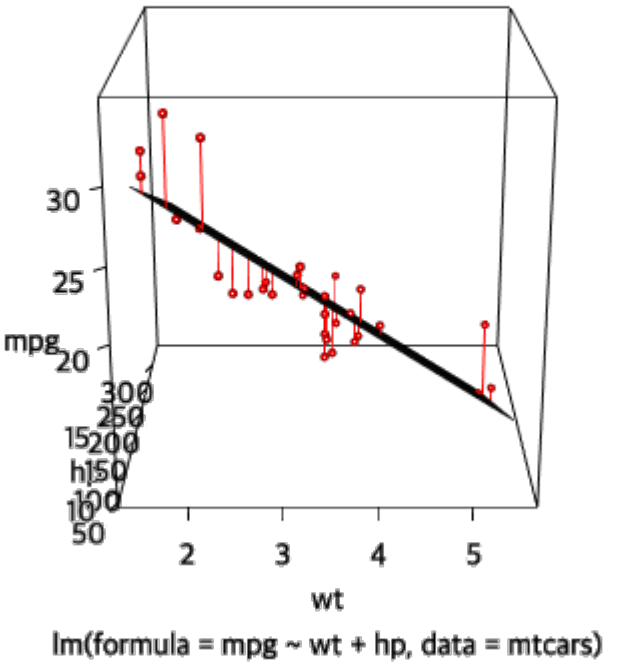
```
ggPredict(fit3,mode=3,colorn=50,show.text = FALSE)
```



```
predict3d(fit3, show.error = TRU
```

```
fit31=lm( mpg ~ wt+hp, data=mtca
predict3d(fit31, show.error = TR
```





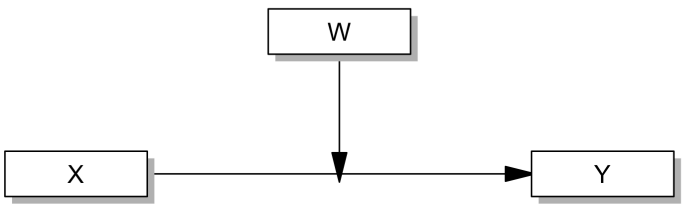
개념적모형

상호작용이 있는 다중회귀분석은 Hayes의 PROCESS macro 모형 1에 해당한다. 다음과 같은 R formula를 사용한다.

> R formula: Y ~ X * W

개념적 모형은 다음과 같다.

```
pmacroModel(1,radx=0.1,rady=0.07)
```



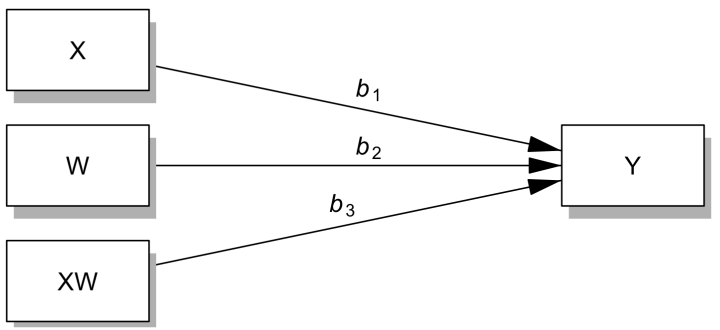
통계적모형

Model syntax는 다음과 같다.

> Model syntax: Y ~ b1 * X + b2 * W + b3 * X:W

통계적 모형은 다음과 같다.

```
statisticalDiagram(1,radx=0.1,rady=0.07)
```



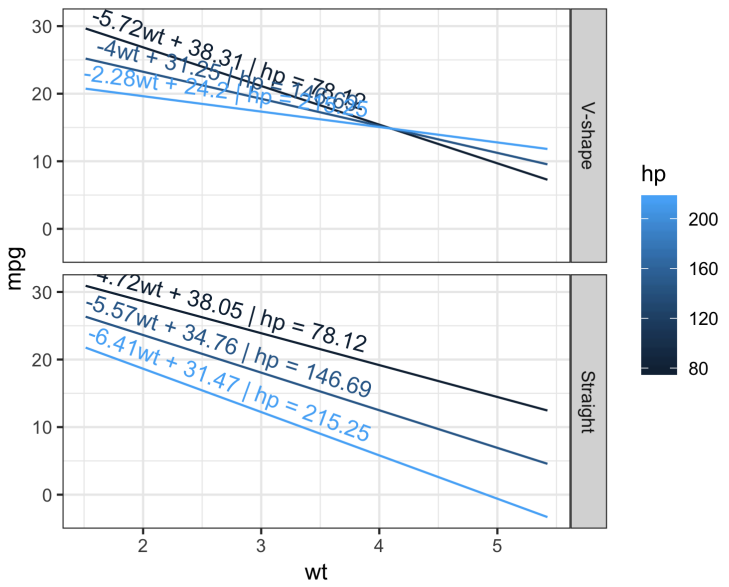
조절된 조절(moderated moderation)

> A * B * C = A + B + C + A:B + B:C + A:C + A:B:C

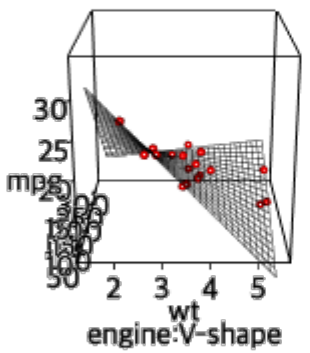
```
fit4=lm(mpg ~ wt * hp * engine,data=mtcars)
summary(fit4)
```

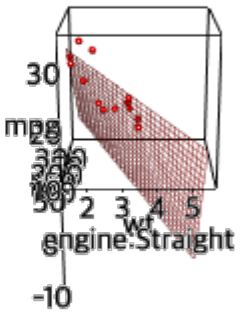
```
Call:
lm(formula = mpg ~ wt * hp * engine, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4392 -1.4404  0.0168  1.3475  3.8171

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          46.34789    9.27508   4.997 4.19e-05 ***
wt                   -7.68255    2.97908  -2.579   0.0165 *
hp                   -0.10291    0.04901  -2.100   0.0464 *
engineStraight       -4.54377   12.64388  -0.359   0.7225
wt:hp                 0.02509    0.01503   1.669   0.1081
wt:engineStraight     3.92911    4.67846   0.840   0.4093
hp:engineStraight     0.05489    0.10581   0.519   0.6086
wt:hp:engineStraight -0.03745    0.03927  -0.954   0.3497
```

```
ggPredict(fit4,show.point=FALSE)
```

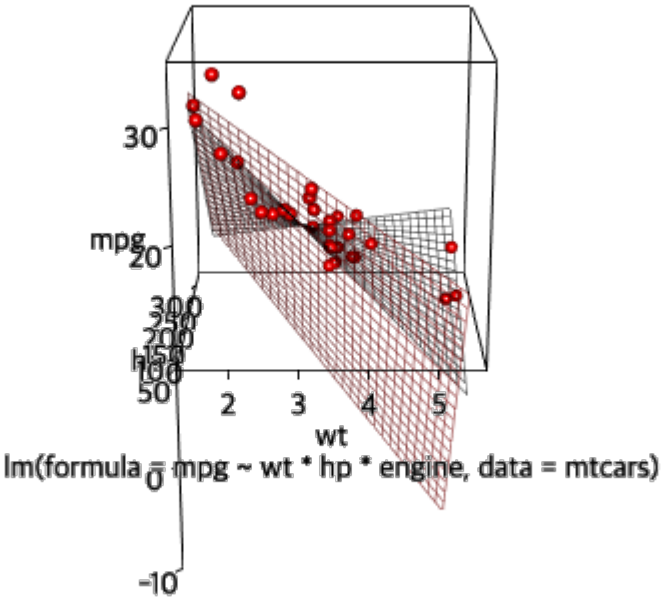


```
predict3d(fit4, radius=4)
```





```
predict3d(fit4, radius=4,overlay=TRUE)
```



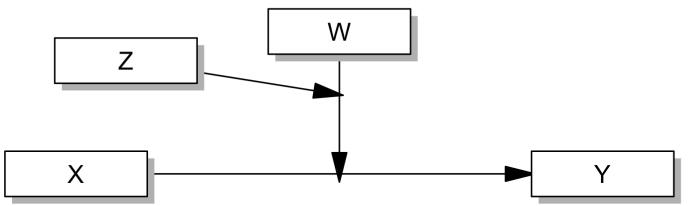
조절된 조절의 개념적모형

설명변수가 3개 있고 모두 상호작용이 있는 모형은 조절된 조절이라고 할 수 있으며 Hayes의 PROCESS macro 모형 3에 해당한다.

> R formula: Y ~ X * W * Z

개념적 모형은 다음과 같다.

```
pmacroModel(3,radx=0.1,rady=0.07)
```



조절된 조절의 통계적모형

Model syntax는 다음과 같다.

> Y ~ b1X + b2 W + b3 Z + b4 X:W + b5 X:Z + b6 W:Z + b7 X:W:Z

```
statisticalDiagram(3)
```