응용경영통계 과제 2019.05.22

20160796 박보성

01 iris 데이터에서 종속변수 sl을 가장 잘 설명하는 선형모형을 구축하시오

어떤 모형이 가장 좋은 모형인지에 대하여 그 생각과 이유를 제시하시오/그 기준에 맞는 가장 좋은 모형을 찾으시오

P-value가 낮고 상대적으로 AIC, BIC가 작은 모형이 좋은 모형이라고 볼 수 있다. 충분히 예측력이 있는 모형이기 때문이다.

```
names(iris)<-c('sl','sw','pl','pw','sp')
ress <- lm(sl~ .,iris)
summary(ress)
```

```
Call:
lm(formula = sl ~ ., data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-0.79424 -0.21874  0.00899  0.20255  0.73103

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.17127    0.27979   7.760 1.43e-12 ***
sw            0.49589    0.08607   5.761 4.87e-08 ***
pl            0.82924    0.06853  12.101  < 2e-16 ***
pw           -0.31516    0.15120  -2.084  0.03889 *
spversicolor -0.72356    0.24017  -3.013  0.00306 **
spvirginica  -1.02350    0.33373  -3.067  0.00258 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3068 on 144 degrees of freedom
Multiple R-squared:  0.8673,    Adjusted R-squared:  0.8627
F-statistic: 188.3 on 5 and 144 DF,  p-value: < 2.2e-16
```

Pw, pl이 상대적으로 예측력이 높은 변수임을 알 수 있다. 따라서 이 두 변수를 이용하여 선형모형을 만들면 좋은 모형을 만들 수 있다.

우선 pl과 관련하여 y 절편이 있는 모형과 없는 모형을 비교해보자.

```
ress1 <- lm(sl~ pl, iris)
ress2 <- lm(sl~-1+pl, iris)
summary(ress1)
summary(ress2)
```

```
> summary(ress1)

Call:
lm(formula = sl ~ pl, data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-1.24675 -0.29657 -0.01515  0.27676  1.00269

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.30660    0.07839   54.94   <2e-16 ***
pl           0.40892    0.01889   21.65   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4071 on 148 degrees of freedom
Multiple R-squared:   0.76,     Adjusted R-squared:  0.7583
F-statistic: 468.6 on 1 and 148 DF,  p-value: < 2.2e-16
```

```
> summary(ress2)

Call:
lm(formula = sl ~ -1 + pl, data = iris)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7933 -0.6583  0.2743  2.7918  4.1813

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
pl  1.34888    0.03692   36.53   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.877 on 149 degrees of freedom
Multiple R-squared:  0.8996,    Adjusted R-squared:  0.8989
F-statistic:  1334 on 1 and 149 DF,  p-value: < 2.2e-16
```

Y 절편이 없는 경우가 R^2는 더 높지만 AIC, BIC를 비교했을 때 훨씬 값이 크기 때문에 Y 절편이 없는 경우가 더 좋은 모형이다.

```
> AIC(ress1, ress2)
      df      AIC
ress1  3 160.0404
ress2  2 617.5056
> BIC(ress1, ress2)
      df      BIC
ress1  3 169.0723
ress2  2 623.5269
```

다양한 모형들을 비교해보았다.

```
rest1= lm(sl~pl,iris)
rest2= lm(sl~pw*pl,iris)
rest3= lm(sl~pw+pl,iris)
rest4= lm(sl~pw:pl,iris)
rest5= lm(sl~I(pl^2)*pw,iris)
rest6= lm(sl~I(pw^2)+pl,iris)
rest7= lm(sl~pl*I(pw^2),iris)
```

```
rest8= lm(sl~pw+I(pl^2),iris)
AIC(rest1, rest2, rest3, rest4, rest5, rest6, rest7, rest8)
BIC(rest1, rest2, rest3, rest4, rest5, rest6, rest7, rest8)
```

```
> AIC(rest1, rest2, rest3, rest4, rest5, rest6, rest7, rest8)
      df      AIC
rest1  3 160.0404
rest2  5 130.6952
rest3  4 158.0468
rest4  3 174.8960
rest5  5 125.1783
rest6  4 161.9298
rest7  5 125.9031
rest8  4 123.2201
> BIC(rest1, rest2, rest3, rest4, rest5, rest6, rest7, rest8)
      df      BIC
rest1  3 169.0723
rest2  5 145.7484
rest3  4 170.0894
rest4  3 183.9279
rest5  5 140.2315
rest6  4 173.9724
rest7  5 140.9563
rest8  4 135.2627
```

rest8= lm(sl~pw+I(pl^2),iris) 가 가장 좋은 모형이라는 결론을 내릴 수 있다.

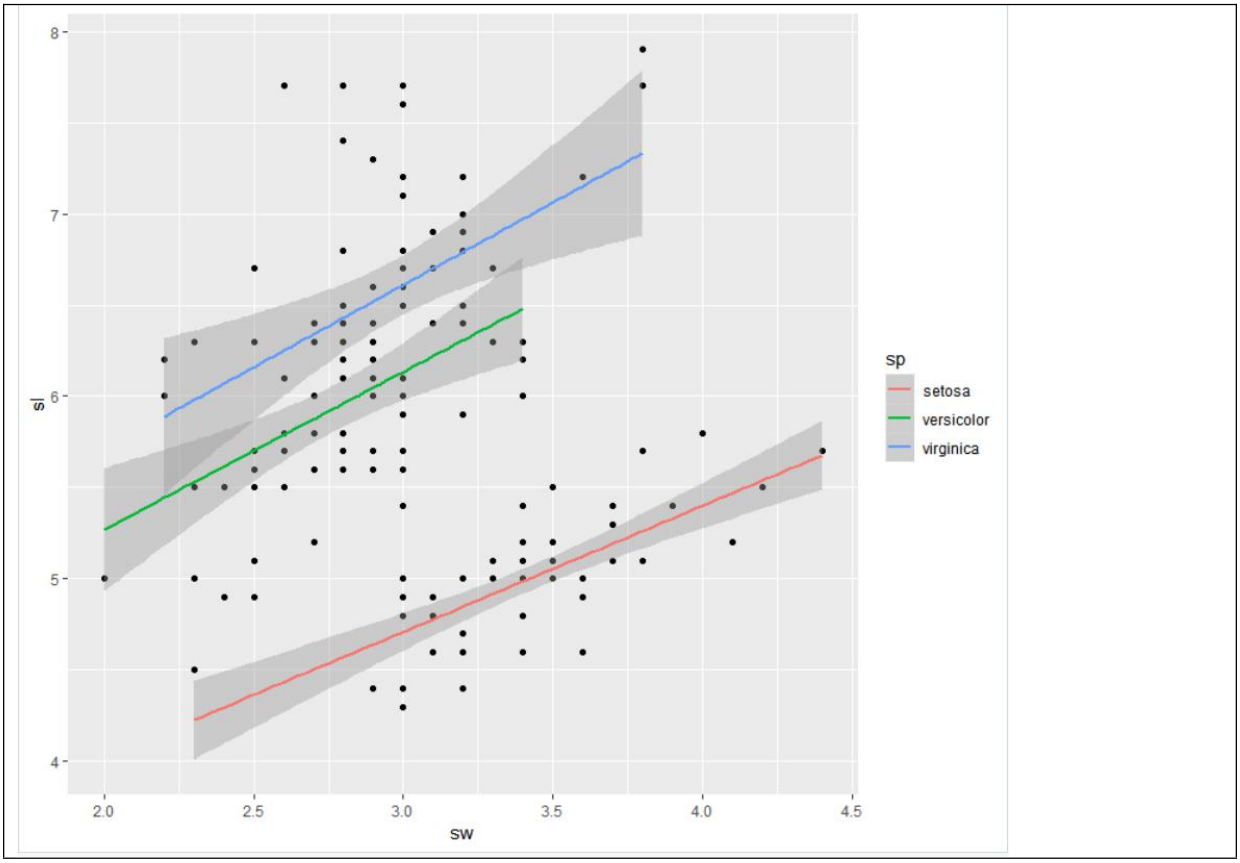
02 iris 데이터에서 종속변수 sl이고 독립변수로 sw, sp를 사용하는 모형에서 교호작용이 있는 경우와 없는 경우의 모형을 구성하여 적합하고 그 결과를 그림으로 표현하고 어떤 모형이 더 좋은 모형이라고 생각하는지 그 근거를 설명하시오

```
res11 = lm(sl~sw+sp,iris)
res12 = lm(sl~sw*sp,iris)
AIC(res11, res12)
```

```
> AIC(res11, res12)
      df      AIC
res11  5 183.9366
res12  7 187.0922
```

교호작용이 없는 경우가 더 좋은 모형이라고 할 수 있다.

```
> library(ggplot2)
> ggplot(iris)+geom_point(aes(x=sw, y=sl))+stat_smooth(method='lm', aes(x=sw, y=sl, col=sp))
> p<-ggplot(data=iris,aes(x=sw,y=sl,colour=sp))+geom_point()+geom_smooth(method="lm")
```



03 mtcars 데이터에서 다음 모형들의 의미와 그 차이를 비교하고 R을 이용하여 적합한 결과 어떤 차이가 나는지 설명하시오

1. Mpg ~ -1+ wt + factor(cyl) #factor 변수가 있고 y 절편이 없는 경우
2. Mpg ~ wt + cyl #y 절편이 있는 경우
3. Mpg ~ -1 + wt + cyl #y 절편이 없는 경우
4. Mpg ~ -1 + wt + I(cyl-6) #cyl가 6씩 줄어든 경우

ANOVA table 비교

```
> res1 <- lm(mpg ~ wt + factor(cyl), mtcars)
> anova(res1)
Analysis of Variance Table

Response: mpg
            Df Sum Sq Mean Sq  F value    Pr(>F)
wt           1 847.73  847.73 129.6650 5.079e-12 ***
factor(cyl)  2  95.26   47.63   7.2856  0.002835 **
Residuals   28 183.06    6.54
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


> res2 <- lm(mpg ~ wt + cyl, mtcars)
> anova(res2)
Analysis of Variance Table

Response: mpg
          Df Sum Sq Mean Sq F value    Pr(>F)
```

```
wt          1 847.73  847.73  128.60 3.535e-12 ***
cyl         1  87.15   87.15   13.22  0.001064 **
Residuals 29 191.17    6.59
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> res3 <- lm(mpg ~ -1 + wt + cyl, mtcars)
> anova(res3)
Analysis of Variance Table

Response: mpg
          Df  Sum Sq Mean Sq F value    Pr(>F)
wt         1 10105.7 10105.7 81.4698 4.704e-10 ***
cyl        1   215.3   215.3  1.7361    0.1976
Residuals 30  3721.3   124.0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> res4 <- lm(mpg ~ -1 + wt + I(cyl-6), mtcars)
> anova(res4)
Analysis of Variance Table

Response: mpg
           Df  Sum Sq Mean Sq F value    Pr(>F)
wt          1 10105.7 10105.7 242.572 6.427e-16 ***
I(cyl - 6)  1  2686.8  2686.8  64.493 5.787e-09 ***
Residuals  30  1249.8    41.7
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA 비교를 통해서 발견할 수 있는 점:

-y 절편이 있는 모형에서 factor 변수의 자유도는 집단수 -1 이다.

-y 절편이 있는 경우와 없는 경우에서 잔차 자유도는 29, 30 으로 나타난다.

Summary Table 비교

```
> summary(res1)

Call:
lm(formula = mpg ~ wt + factor(cyl), data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5890 -1.2357 -0.5159  1.3845  5.7915

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   33.9908     1.8878  18.006  < 2e-16 ***
wt            -3.2056     0.7539  -4.252 0.000213 ***
factor(cyl)6  -4.2556     1.3861  -3.070 0.004718 **
factor(cyl)8  -6.0709     1.6523  -3.674 0.000999 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.557 on 28 degrees of freedom
Multiple R-squared:  0.8374,   Adjusted R-squared:   0.82
F-statistic: 48.08 on 3 and 28 DF,  p-value: 3.594e-11

> summary(res2)
```

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

Residuals:
    Min      1Q  Median      3Q     Max
-4.2893 -1.5512 -0.4684  1.5743  6.1004

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  39.6863     1.7150  23.141  < 2e-16 ***
wt           -3.1910     0.7569  -4.216 0.000222 ***
cyl          -1.5078     0.4147  -3.636 0.001064 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.568 on 29 degrees of freedom
Multiple R-squared:  0.8302,   Adjusted R-squared:  0.8185
F-statistic: 70.91 on 2 and 29 DF,  p-value: 6.809e-12

> summary(res3)

Call:
lm(formula = mpg ~ -1 + wt + cyl, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-13.466  -6.181   1.476  10.597  22.997

Coefficients:
    Estimate Std. Error t value Pr(>|t|)
wt     1.174      3.180   0.369    0.715
cyl    2.187      1.660   1.318    0.198

Residual standard error: 11.14 on 30 degrees of freedom
Multiple R-squared:  0.735,    Adjusted R-squared:  0.7173
F-statistic:  41.6 on 2 and 30 DF,  p-value: 2.232e-09

> summary(res4)

Call:
lm(formula = mpg ~ -1 + wt + I(cyl - 6), data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-12.422  -3.494   2.532   4.832  11.510

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
wt            6.2284     0.3592  17.339  < 2e-16 ***
I(cyl - 6)   -5.4805     0.6824  -8.031 5.79e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.455 on 30 degrees of freedom
Multiple R-squared:  0.911,    Adjusted R-squared:  0.9051
F-statistic: 153.5 on 2 and 30 DF,  p-value: < 2.2e-16
```

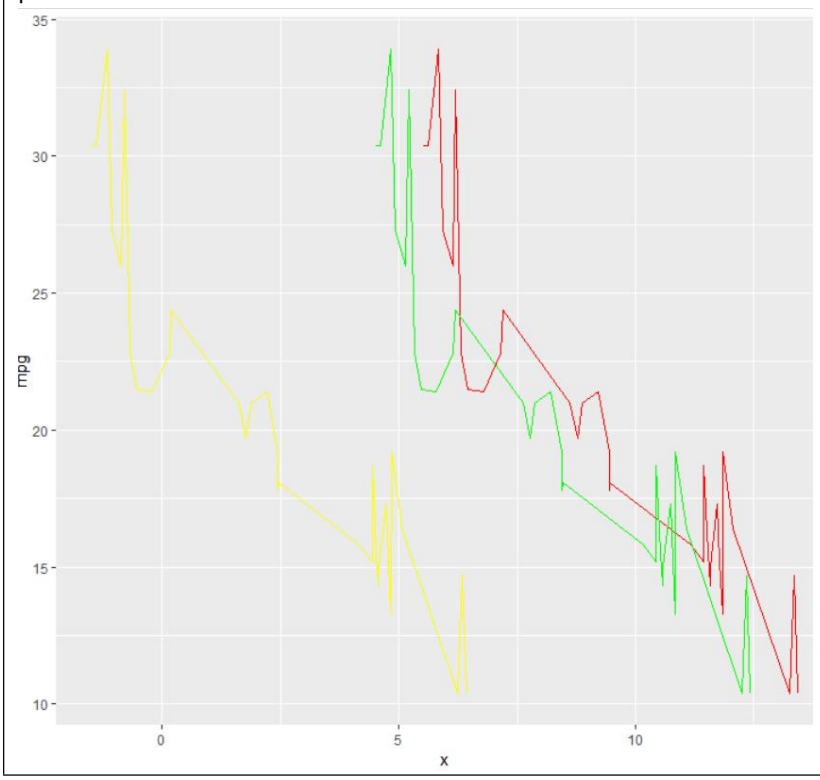
- 비교했을 때 res4가 가장R^2가 높은 값이 나왔다.
- Factor 변수가 있는 경우(res1) 와 없는 경우(res2)를 비교했을 때는 factor변수가 있는 경우가 더 좋은 모형이다. (p value더 낮고 R^2높다)
- Res2, res3 비교했을 때는 Y 절편이 없는 모형보다 있는 모형이 더 좋다.

```
> AIC(res1, res2, res3, res4)
     df      AIC
res1  5 156.6223
res2  4 156.0101
res3  3 249.0067
res4  3 214.0926
> BIC(res1, res2, res3, res4)
     df      BIC
res1  5 163.9510
res2  4 161.8730
res3  3 253.4040
res4  3 218.4899
```

- AIC, BIC를 비교해 보았을 때는 res2 가 가장 낮은 값이 나왔다.

그래프로 그려서 비교해보기

```
ff <- factor(mtcars$cyl)
p = ggplot() +
  #geom_line(data = mtcars, aes(x = -1+ wt + ff, y = mpg), color = "blue") +
  geom_line(data = mtcars, aes(x = wt + cyl, y = mpg), color = "red")  +
  geom_line(data = mtcars, aes(x = -1 + wt + cyl, y = mpg), color = "green")  +
  geom_line(data = mtcars, aes(x =  -1 + wt + I(cyl-6), y = mpg), color = "yellow") +
  xlab('x') +
  ylab('mpg')

p
```

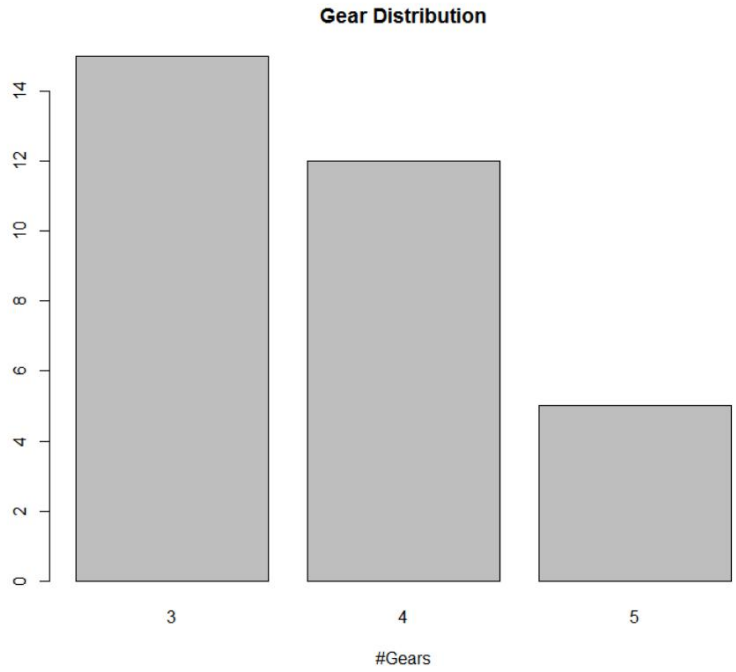


04 mtcars데이터에서, mpg 변수를 종속변수로 하는 가장 좋은 모형을 구상하려 한다.

1) gear변수를 독립변수 중 하나로 포함하는 모형

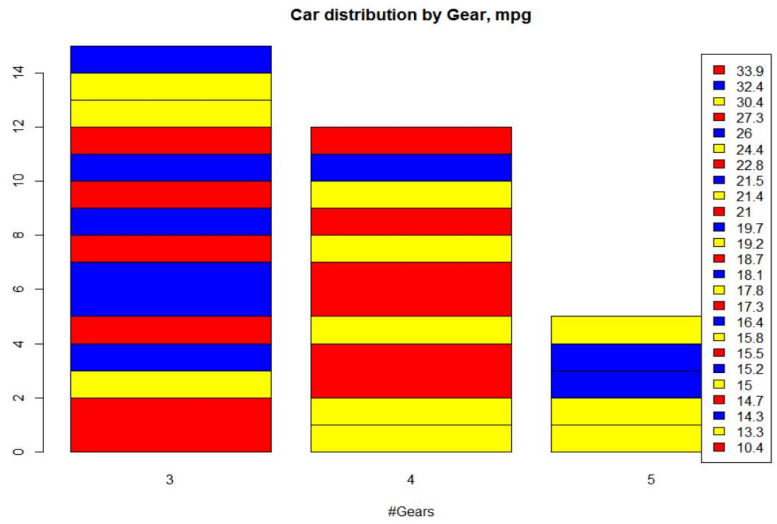
gear의 distribution

```
barplot(table(mtcars$gear),main = "Gear Distribution", xlab = "#Gears")
```



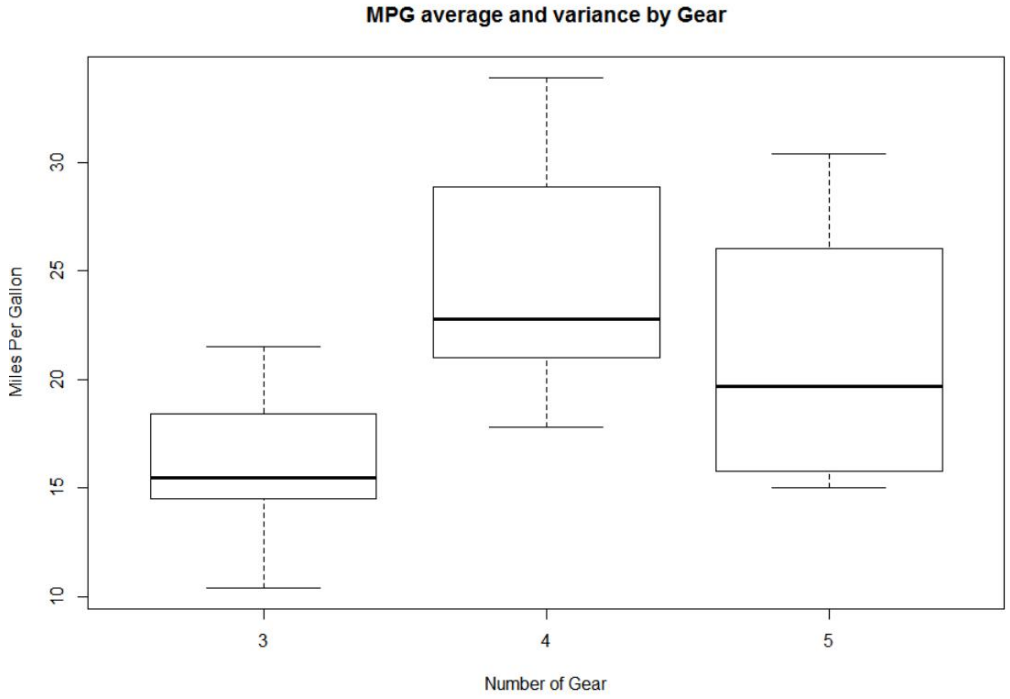
Car Distribution by gear, mpg

```
Counts <- table(mtcars$mpg, mtcars$gear)
barplot(Counts, main = "Car distribution by Gear, mpg", xlab = "#Gears", col  = c("Red", "
Yellow", "Blue"), legend = rownames(Counts))
```




```
boxplot(mpg~gear,data=mtcars, main="MPG average and variance by Gear",
```

```
      xlab="Number of Gear", ylab="Miles Per Gallon")
```



Mpg와 전체 변수들의 관계 알아보기

```
allmt = lm(data = mtcars, mpg ~ .)
summary(allmt)
```

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

Residuals:
    Min      1Q  Median      3Q     Max
-3.4506 -1.6044 -0.1196  1.2193  4.6271

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.30337   18.71788   0.657   0.5181
cyl         -0.11144    1.04502  -0.107   0.9161
disp         0.01334    0.01786   0.747   0.4635
hp          -0.02148    0.02177  -0.987   0.3350
drat         0.78711    1.63537   0.481   0.6353
wt          -3.71530    1.89441  -1.961   0.0633 .
qsec         0.82104    0.73084   1.123   0.2739
vs           0.31776    2.10451   0.151   0.8814
am           2.52023    2.05665   1.225   0.2340
gear         0.65541    1.49326   0.439   0.6652
carb        -0.19942    0.82875  -0.241   0.8122
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.65 on 21 degrees of freedom
Multiple R-squared:  0.869,     Adjusted R-squared:  0.8066
F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

Wt가 가장 mpg에 따라서 움직이는 변수라는 것을 알 수 있다.

그 다음 amManual이 어느정도 설명력이 있다고 판단하여 mpg를 wt, gear, amManual로 설명해보기로 결정했다.
설명변수가 늘어나면 당연히 R^2값이 늘어나기 때문에 amManual을 포함했을 때, 포함하지 않았을 때의 AIC 값을 비교해보았다. 포함하지 않았을 때가 더 낮게 나왔다.

```
nmd1 = lm(data = mtcars, mpg ~ wt + gear + am)
        summary(nmd)


nnn = lm(data = mtcars, mpg ~ wt + gear)
        summary(nnn)
AIC(nmd1, nnn)
> AIC(nmd1, nnn)
     df      AIC
nmd1  5 169.8073
nnn   4 167.8984
```

y 절편이 있는 경우에는

```
nmd = lm(data = mtcars, mpg ~ wt + gear)

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

Residuals:
    Min      1Q  Median      3Q     Max
-4.1304 -2.3061 -0.2932  1.4409  6.8296

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  38.9157     5.0974   7.634 2.04e-08 ***
wt           -5.4850     0.6987  -7.851 1.17e-08 ***
gear         -0.3196     0.9265  -0.345    0.733
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.092 on 29 degrees of freedom
Multiple R-squared:  0.7538,    Adjusted R-squared:  0.7369
F-statistic: 44.41 on 2 and 29 DF,  p-value: 1.488e-09
```

모형이 가장 좋은 모형이다.

y 절편이 없고 식에 변형을 조금 주었을 경우 훨씬 더 좋은 모형이 된다.

```
nnn1 = lm(data = mtcars, mpg ~ -1+ I(wt^2) + gear)
summary(nnn1)
```

```
Call:
lm(formula = mpg ~ -1 + I(wt^2) + gear, data = mtcars)

Residuals:
     Min       1Q   Median       3Q      Max
-12.1894  -1.3621   0.3398   3.2190  10.5849

Coefficients:
        Estimate Std. Error t value Pr(>|t|)
I(wt^2)  -0.2290     0.1123  -2.039   0.0504 .
gear      6.0215     0.3925  15.340 9.64e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.145 on 30 degrees of freedom
Multiple R-squared:  0.9435,    Adjusted R-squared:  0.9397
F-statistic: 250.3 on 2 and 30 DF,  p-value: < 2.2e-16
```

2) gear변수를 ordered변수로 변환해서 포함하는 모형

명목변수가 아닌 순서변수로 gear를 포함시키는 모형이다. Polynomial contrast로 변수를 대입하게 된다.

```
onmd = lm(data = mtcars, mpg ~ -1+I(wt^2) + ordered(gear))
summary(onmd)
Call:
lm(formula = mpg ~ -1 + I(wt^2) + ordered(gear), data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0643 -2.8524 -0.1987  2.1366  7.1084

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
I(wt^2)          -0.5871     0.1181  -4.971 3.00e-05 ***
ordered(gear)3   25.3824     2.0725  12.247 9.20e-13 ***
ordered(gear)4   28.7684     1.3199  21.795  < 2e-16 ***
ordered(gear)5   25.7638     1.7936  14.364 1.92e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.492 on 28 degrees of freedom
Multiple R-squared:  0.9757,    Adjusted R-squared:  0.9722
F-statistic: 280.9 on 4 and 28 DF,  p-value: < 2.2e-16
```

R^2값은 더 높아지고 P-value는 더 낮아졌다.

3) 결과를 비교하여 설명하기

```
> AIC(onmd, nnn1)
     df      AIC
onmd  5 176.5736
nnn1  3 199.5769
> BIC(onmd, nnn1)
     df      BIC
onmd  5 183.9023
nnn1  3 203.9741
```

AIC값과 BIC 값을 비교했을 때 ordered인 경우가 더 낮게 나온다.

따라서 ordered(gear)를 포함한 경우가 더 좋은 모형이라고 할 수 있다.

gear가 아니라 wt를 ordered로 넣었을 경우에도 기존보다 훨씬 좋은 모형으로 변했다.

```
onmd1 = lm(data = mtcars, mpg ~ -1+ordered(wt^2) + gear)
summary(onmd1)
```

```
                       Estimate Std. Error t value Pr(>|t|)
ordered(wt^2)2.289169   29.3375     2.4225  12.111  0.00675 **
ordered(wt^2)2.608225   29.5500     1.9898  14.851  0.00450 **
ordered(wt^2)3.367225   33.0500     1.9898  16.610  0.00361 **
ordered(wt^2)3.744225   26.4500     1.9898  13.293  0.00561 **
ordered(wt^2)4.5796     24.9375     2.4225  10.294  0.00930 **
ordered(wt^2)4.84       31.5500     1.9898  15.856  0.00395 **
ordered(wt^2)5.3824     21.9500     1.9898  11.031  0.00812 **
ordered(wt^2)6.076225   20.8625     1.5731  13.262  0.00564 **
ordered(wt^2)6.8644     20.1500     1.9898  10.127  0.00961 **
ordered(wt^2)7.6729     18.6375     2.4225   7.694  0.01648 *
ordered(wt^2)7.7284     20.5500     1.9898  10.328  0.00925 **
ordered(wt^2)8.265625   20.1500     1.9898  10.127  0.00961 **
ordered(wt^2)9.9225     21.9500     1.9898  11.031  0.00812 **
ordered(wt^2)10.0489    14.7375     2.4225   6.084  0.02597 *
ordered(wt^2)10.1761    23.5500     1.9898  11.835  0.00706 **
ordered(wt^2)10.336225  20.7625     1.5731  13.199  0.00569 **
ordered(wt^2)11.799225  14.5625     1.5731   9.257  0.01147 *
ordered(wt^2)11.8336    17.7875     1.7436  10.201  0.00947 **
ordered(wt^2)11.9716    17.4625     1.5731  11.101  0.00802 **
ordered(wt^2)12.3904    14.8625     1.5731   9.448  0.01102 *
ordered(wt^2)12.7449    13.8000     1.9174   7.197  0.01876 *
ordered(wt^2)13.9129    16.6625     1.5731  10.592  0.00880 **
ordered(wt^2)14.2884    14.5625     1.5731   9.257  0.01147 *
ordered(wt^2)14.7456    12.6625     1.5731   8.049  0.01509 *
ordered(wt^2)14.784025  18.5625     1.5731  11.800  0.00711 **
ordered(wt^2)16.5649    15.7625     1.5731  10.020  0.00981 **
ordered(wt^2)27.5625     9.7625     1.5731   6.206  0.02500 *
ordered(wt^2)28.569025  14.0625     1.5731   8.939  0.01228 *
ordered(wt^2)29.419776   9.7625     1.5731   6.206  0.02500 *
gear                     0.2125     0.4606   0.461  0.68983
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7521 on 2 degrees of freedom
Multiple R-squared:  0.9999,    Adjusted R-squared:  0.9987
F-statistic: 827.5 on 30 and 2 DF,  p-value: 0.001208
```

p-value와 R^2값 모두 굉장히 유의미하게 나왔으며

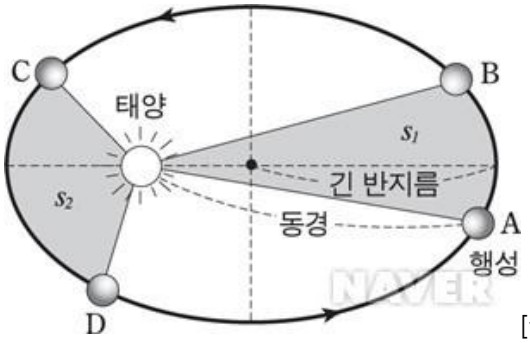
```
> AIC(nnn1,onmd, onmd1)
      df       AIC
nnn1   3 199.57689
onmd   5 176.57363
onmd1 31  45.85486
> BIC(nnn1,onmd, onmd1)
      df       BIC
nnn1   3 203.97410
onmd   5 183.90231
onmd1 31  91.29267
```

AIC, BIC값 또한 굉장히 낮게 나왔다.

05 케플러의 제3 법칙과 회귀분석

독일 천문학자인 Kepler가 덴마크 천문학자 Brahe의 관측결과로부터 얻은 행성운동에 대한 법칙을

만들었다. 그 중에 제 3 법칙은 행성의 공전주기 T의 제곱이 타원 궤도의 긴 반지름인 R의 3제곱에 비례한다는 것으로 식으로는 $T^2=kR^3$ (k: 비례상수) 로 표현할 수 있다.



[출처: Wikipedia, naver사전]

선형회귀를 알아보기 위한 input data는 다음과 같이 정리할 수 있다.

Planet	Semi-Major Axis x	Period P
Mercury	0.39	0.24
Venus	0.72	0.61
Earth	1.00	1.00
Mars	1.52	1.88
Jupiter	5.20	11.86
Saturn	9.54	29.46
Uranus	19.19	84.01
Neptune	30.06	164.79
Pluto	39.53	248.54

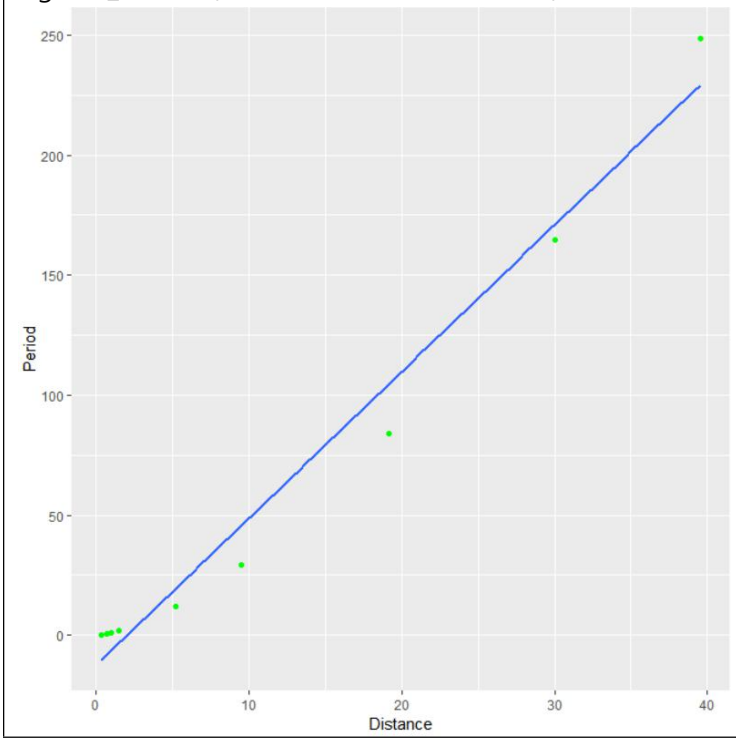
Planets, Distance, Period를 넣어서 Kepler를 만든다.

```
library(ggplot2)
Planets <- c('Mercury', 'Venus', 'Earth', 'Mars', 'Jupiter', 'Saturn', 'Uranus', 'Neptune', 'Pluto')
Distance <- c(0.39, 0.72 ,1.00, 1.52, 5.20, 9.54, 19.19, 30.06, 39.53)
Period <- c(0.24, 0.61, 1.00, 1.88, 11.86, 29.46, 84.01, 164.79, 248.54)

kepler = data.frame(Planets, Distance, Period)
kepler
```

```
> kepler
  Planets Distance Period
1 Mercury     0.39   0.24
2   Venus     0.72   0.61
3   Earth     1.00   1.00
4    Mars     1.52   1.88
5 Jupiter     5.20  11.86
6  Saturn     9.54  29.46
7  Uranus    19.19  84.01
8 Neptune    30.06 164.79
9   Pluto    39.53 248.54
>
```

Kepler에서 Distance, Period를 비교했을 때

```
a= lm( Period~Distance)
summary(a)

ggplot(data = kepler, aes(x = Distance, y = Period)) +
  geom_point(color='green') +
  geom_smooth(method = "lm", se = FALSE)
```



다음과 같은 선형 관계가 있음을 확인할 수 있다.

*T^2=kR^3*을 확인하기 위해서는 상용로그를 취해서 값을 구했을 때 기울기를 확인하면 된다.

T=R ^(3/2) 이기 때문에 loga =0, a=1이 되고 n = 1.5=3/2가 되는 것을 알 수 있다. 이를 증명하면 케플러 3법칙이 증명된다.

```
b= lm(log(Period, base=10)~ log(Distance, base=10))
summary(b)
```

```
Call:
lm(formula = log(Period, base = 10) ~ log(Distance, base = 10))

Residuals:
       Min         1Q     Median         3Q        Max
-0.0042597 -0.0005731  0.0000888  0.0010836  0.0026899

Coefficients:
                          Estimate Std. Error  t value Pr(>|t|)
(Intercept)              -0.0015524  0.0009327   -1.664     0.14
log(Distance, base = 10)  1.5014025  0.0009847 1524.678   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.002085 on 7 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 2.325e+06 on 1 and 7 DF,  p-value: < 2.2e-16
```

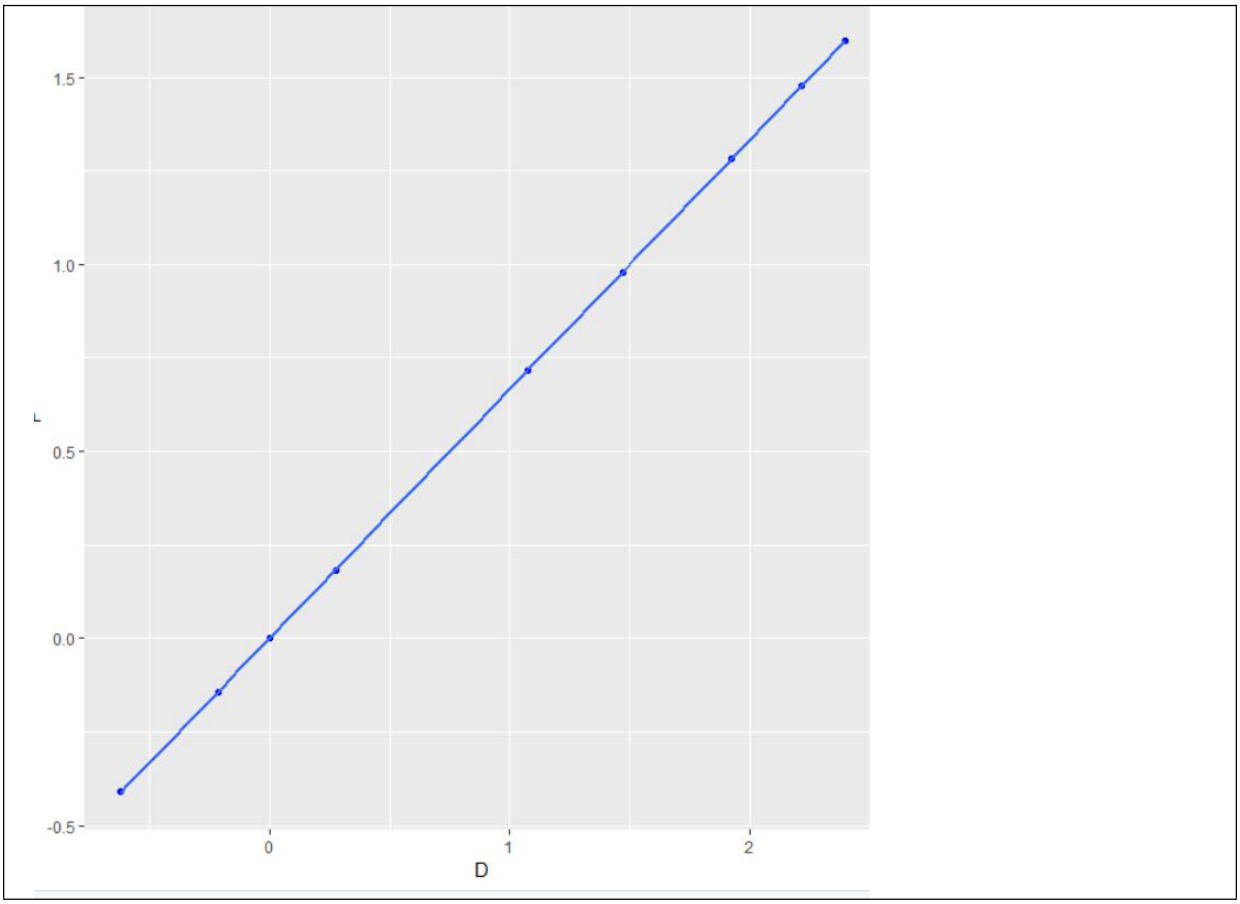
기울기가 1.5로 확인되었다.

그래프로 확인해보면 다음과 같다.

```
b= lm(log(Period, base=10)~ log(Distance, base=10))
summary(b)
```

```
Call:
lm(formula = log(Period, base = 10) ~ log(Distance, base = 10))

Residuals:
       Min         1Q     Median         3Q        Max
-0.0042597 -0.0005731  0.0000888  0.0010836  0.0026899

Coefficients:
                          Estimate Std. Error  t value Pr(>|t|)
(Intercept)              -0.0015524  0.0009327   -1.664     0.14
log(Distance, base = 10)  1.5014025  0.0009847 1524.678   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.002085 on 7 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 2.325e+06 on 1 and 7 DF,  p-value: < 2.2e-16

> D = (log(Period, base=10))
> P = (log(Distance, base=10))
> data.frame(Planets, log(Distance, base=10),log(Period, base=10))
  Planets log.Distance..base...10. log.Period..base...10.
1 Mercury               -0.4089354             -0.6197888
2   Venus               -0.1426675             -0.2146702
3   Earth                0.0000000              0.0000000
4    Mars                0.1818436              0.2741578
5 Jupiter                0.7160033              1.0740847
6  Saturn                0.9795484              1.4692327
7  Uranus                1.2830750              1.9243310
8 Neptune                1.4779890              2.2169309
9   Pluto                1.5969268              2.3953963
>
```

```
ggplot(data = kepler, aes(x = D, y = P)) +
  geom_point(color='blue') +
  geom_smooth(method = "lm", se = FALSE)
```



참고자료

http://ime.math.arizona.edu/g-teams/Profiles/VP/KeplersLawsRegression.pdf