다중회귀분석

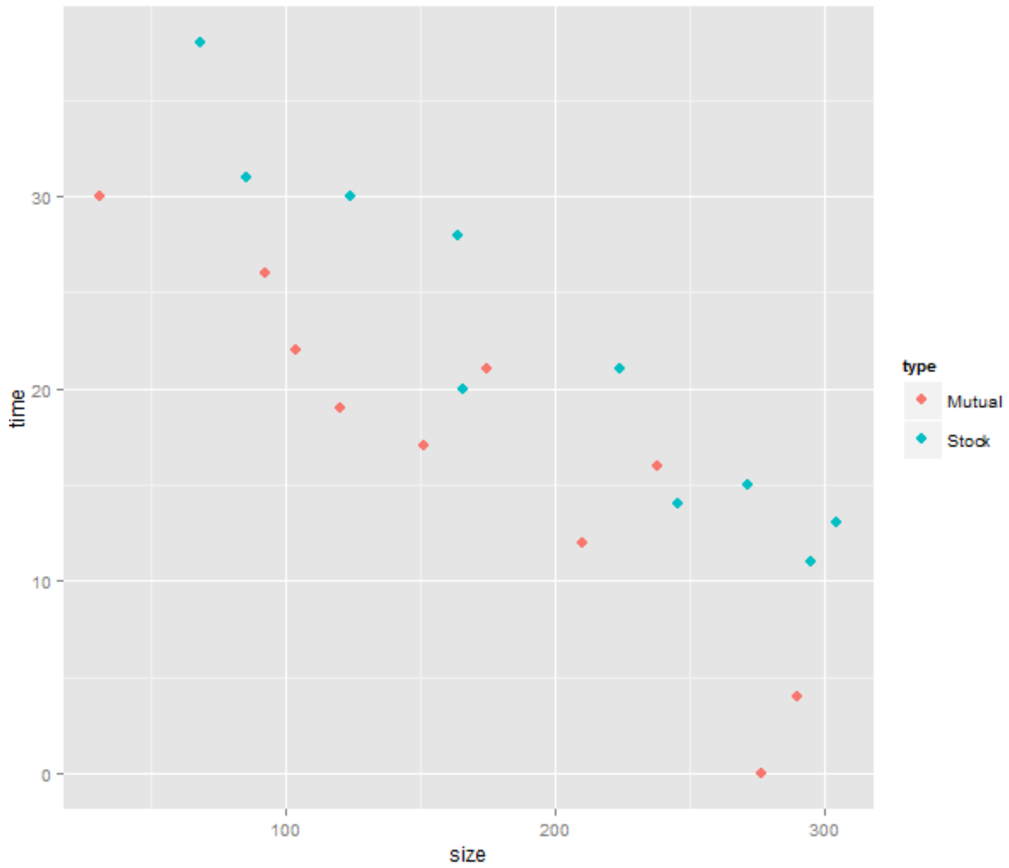
범주형 설명변수

예: Innovation in Insurance Industry

- 보험업계의 혁신에 대한 연구
- 혁신을 받아들이는데 걸리는 시간이 회사 규모와 유형에 따라 달라지는가?
  - Y: 혁신을 받아들이는데 까지 걸리는 기간을 월 단위로 측정
  - X1: 회사의 자산규모
  - X2: 회사 유형 (stock, mutual)

```
> insurance
   time size   type
1    17  151 Mutual
2    26   92 Mutual
3    21  175 Mutual
4    30   31 Mutual
5    22  104 Mutual
6     0  277 Mutual
7    12  210 Mutual
8    19  120 Mutual
9     4  290 Mutual
10   16  238 Mutual
11   28  164  Stock
12   15  272  Stock
13   11  295  Stock
14   38   68  Stock
15   31   85  Stock
16   21  224  Stock
17   20  166  Stock
18   13  305  Stock
19   30  124  Stock
20   14  246  Stock
```

```r
library(ggplot2)
ggplot(insurance, aes(y=time,x=size,color=type))+
  geom_point(size=3)
```



회귀모형

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

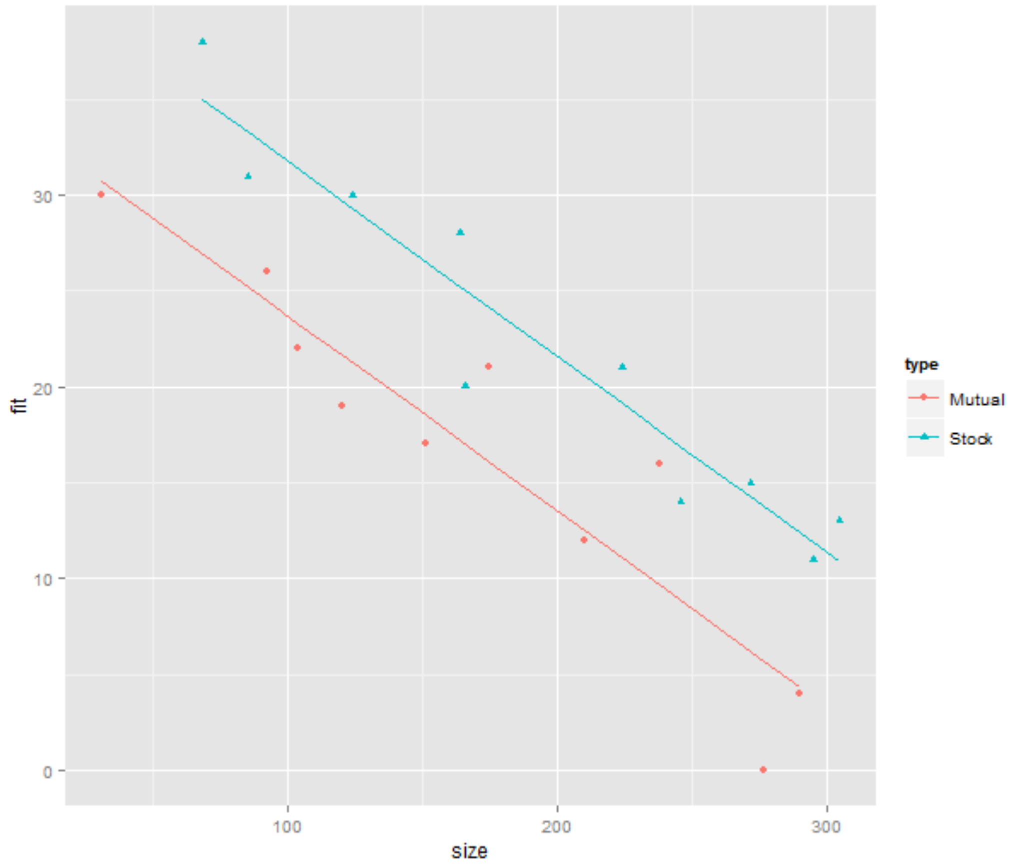
- X_1: size
- X_2: type (=1 if Stock, 0 if Mutual)

- 각 집단의 회귀식
  - Mutual firms: $E(y) = \beta_0 + \beta_1 X_1$
  - Stock firms: $E(y) = (\beta_0 + \beta_2) + \beta_1 X_1$
- β_2의 해석
  - Mutual firm에 비해 Stock firm의 회귀식이 얼만큼 높은가?
  - T-test 를 통해 $H_0: \beta_2 = 0$ 를 검정

```
> summary(insurance)
      time            size            type
 Min.   : 0.00   Min.   : 31.0   Mutual:10
 1st Qu.:13.75   1st Qu.:116.0   Stock :10
 Median :19.50   Median :170.5
 Mean   :19.40   Mean   :181.8
 3rd Qu.:26.50   3rd Qu.:252.5
 Max.   :38.00   Max.   :305.0
> model.matrix(time~size+type,data=insurance)
   (Intercept) size typeStock
1            1  151         0
2            1   92         0
3            1  175         0
4            1   31         0
5            1  104         0
6            1  277         0
7            1  210         0
8            1  120         0
9            1  290         0
10           1  238         0
11           1  164         1
12           1  272         1
13           1  295         1
14           1   68         1
15           1   85         1
16           1  224         1
17           1  166         1
18           1  305         1
19           1  124         1
20           1  246         1
attr(,"assign")
[1] 0 1 2
attr(,"contrasts")
attr(,"contrasts")$type
[1] "contr.treatment"
```

```
> model1=lm(time~.,insurance)
> summary(model1)

Call:
lm(formula = time ~ ., data = insurance)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.6915 -1.7036 -0.4385  1.9210  6.3406 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 33.874069   1.813858  18.675 9.15e-13 ***
size        -0.101742   0.008891 -11.443 2.07e-09 ***
typeStock    8.055469   1.459106   5.521 3.74e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.221 on 17 degrees of freedom
Multiple R-squared:  0.8951,	Adjusted R-squared:  0.8827 
F-statistic:  72.5 on 2 and 17 DF,  p-value: 4.765e-09
```

```
insurance$fit=model1$fitted
ggplot(insurance,aes(y=fit,x=size,group=type,color=type))+
  geom_line()+
  geom_point(aes(y=time,x=size,shape=type))
```



더미변수 생성: Reference level 조정

- 만일 mutual firm이 기준이 아니라 stock firm을 기준으로 비교하고 싶다면?

- X_2: type (=0 if Stock, 1 if Mutual)

- 각 집단의 회귀식
  - Mutual firms: $E(y) = (\beta_0 + \beta_2) + \beta_1 X_1$
  - Stock firms: $E(y) = \beta_0 + \beta_1 X_1$

```
> insurance$type=relevel(insurance$type,ref="Stock")
> model2=lm(time~size+type,insurance)
> summary(model2)

Call:
lm(formula = time ~ size + type, data = insurance)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.6915 -1.7036 -0.4385  1.9210  6.3406 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 41.929538   2.010101  20.859 1.50e-13 ***
size        -0.101742   0.008891 -11.443 2.07e-09 ***
typeMutual  -8.055469   1.459106  -5.521 3.74e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.221 on 17 degrees of freedom
Multiple R-squared:  0.8951, Adjusted R-squared:  0.8827 
F-statistic:  72.5 on 2 and 17 DF,  p-value: 4.765e-09
```

```
> model.matrix(time~size+type,data=insurance)
   (Intercept) size type1
1            1  151     1
2            1   92     1
3            1  175     1
4            1   31     1
5            1  104     1
6            1  277     1
7            1  210     1
8            1  120     1
9            1  290     1
10           1  238     1
11           1  164     0
12           1  272     0
13           1  295     0
14           1   68     0
15           1   85     0
16           1  224     0
17           1  166     0
18           1  305     0
```

더미변수 생성: Effect Coding

- 만일 특정한 집단을 기준으로 하는 것이 아니라 전체의 평균을 기준으로 비교하고 싶다면?

- X_2: type (=-1 if Stock, 1 if Mutual)

- 각 집단의 회귀식
  - Mutual firms: $E(y) = (\beta_0 + \beta_2) + \beta_1 X_1$
  - Stock firms: $E(y) = (\beta_0 - \beta_2) + \beta_1 X_1$
- β_0: 평균 y 절편
- $+\beta_2$: 전체 평균에 비해 mutual firm의 절편이 얼만큼 큰가?
- $-\beta_2$: 전체 평균에 비해 stock firm의 절편이 얼만큼 큰가?

```
> model.matrix(time~size+type,data=insurance,
+              contrasts = list(type = contr.sum))
   (Intercept) size type1
1            1  151     1
2            1   92     1
3            1  175     1
4            1   31     1
5            1  104     1
6            1  277     1
7            1  210     1
8            1  120     1
9            1  290     1
10           1  238     1
11           1  164    -1
12           1  272    -1
13           1  295    -1
14           1   68    -1
15           1   85    -1
16           1  224    -1
17           1  166    -1
18           1  305    -1
19           1  124    -1
20           1  246    -1
attr(,"assign")
[1] 0 1 2
attr(,"contrasts")
attr(,"contrasts")$type
       [,1]
Mutual    1
Stock    -1
```

```
> model3=lm(time~size+type,insurance,contrasts = list(type = contr.sum))
> summary(model3)

Call:
lm(formula = time ~ size + type, data = insurance, contrasts = list(type = contr.sum))

Residuals:
    Min      1Q  Median      3Q     Max 
-5.6915 -1.7036 -0.4385  1.9210  6.3406 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 37.901804   1.770041  21.413 9.78e-14 ***
size        -0.101742   0.008891 -11.443 2.07e-09 ***
type1       -4.027735   0.729553  -5.521 3.74e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.221 on 17 degrees of freedom
Multiple R-squared:  0.8951,	Adjusted R-squared:  0.8827 
F-statistic:  72.5 on 2 and 17 DF,  p-value: 4.765e-09
```

두 개 이상의 level을 가진 범주형 변수

- 예: County demographic information (CDI)
  - 미국의 가장 인구가 많은 440개 county의 자료
  - Crime: 범죄 발생수 (y)
  - Pop: 인구 (X1)
  - Region: 지역 (1=NE, 2=NC, 3=S, 4=W)
- 범죄발생 건수가 인구, 실업률과 지역에 따라 어떻게 달라지는가?

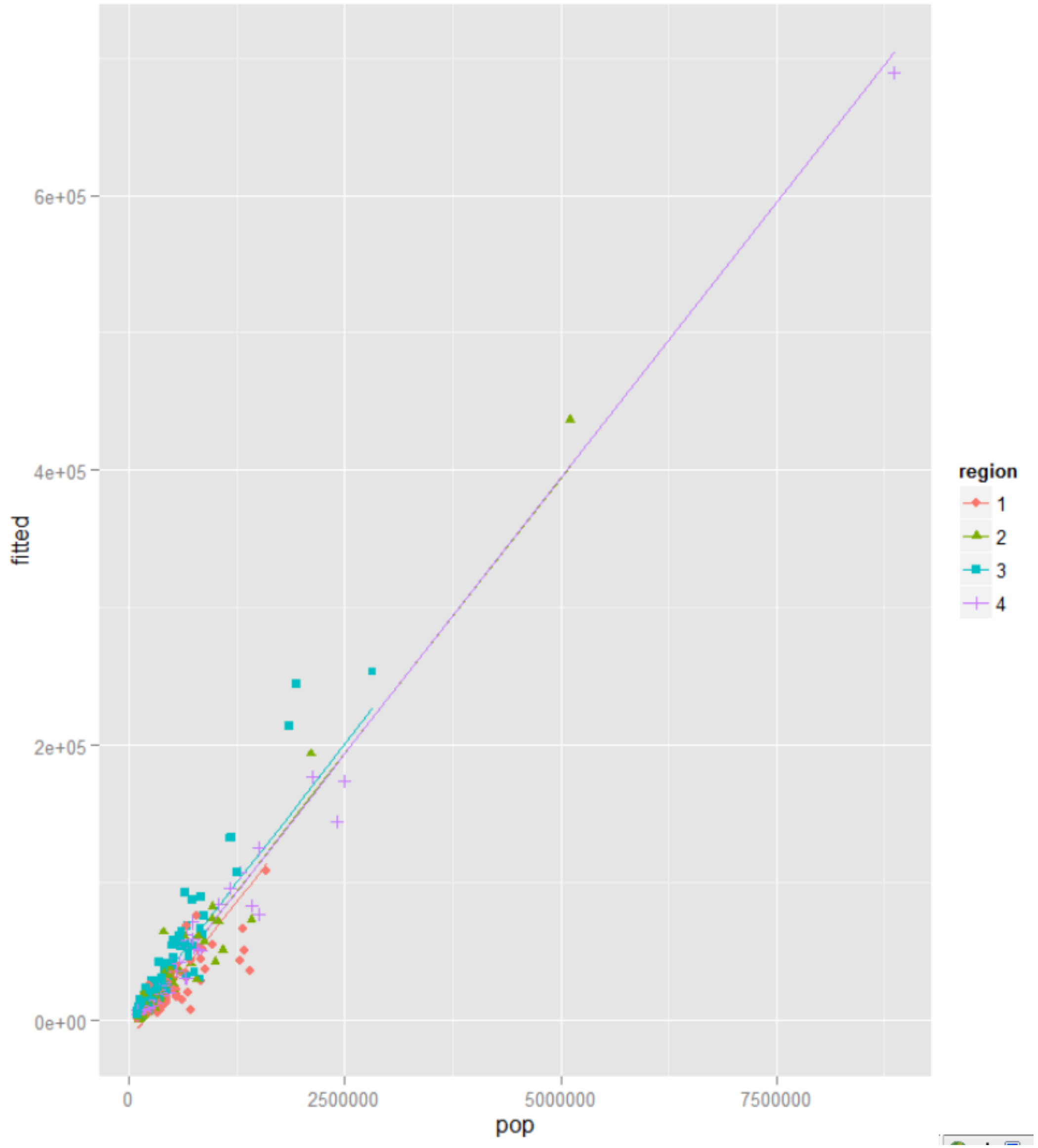
$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_{2,2} + \beta_3 X_{2,3} + \beta_4 X_{2,4} + \epsilon$$

- $X_{2,2}$=1 if region=2, 0 if otherwise
- $X_{2,3}$=1 if region=3, 0 if otherwise
- $X_{2,4}$=1 if region=4, 0 if otherwise

- NE region: $E(y) = \beta_0 + \beta_1 X_1$
- NC region: $E(y) = (\beta_0 + \beta_2) + \beta_1 X_1$
- S region: $E(y) = (\beta_0 + \beta_3) + \beta_1 X_1$
- W region: $E(y) = (\beta_0 + \beta_4) + \beta_1 X_1$

```
> CDI2=CDI[-6,]
> model1=lm(crime~pop+region,CDI2)
> summary(model1)

Call:
lm(formula = crime ~ pop + region, data = CDI2)

Residuals:
   Min     1Q Median     3Q    Max 
-62597  -4366    230   4326  88959 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.357e+04  1.221e+03 -11.111  < 2e-16 ***
pop          8.008e-02  9.561e-04  83.754  < 2e-16 ***
region2      7.626e+03  1.627e+03   4.687 3.71e-06 ***
region3      1.421e+04  1.509e+03   9.420  < 2e-16 ***
region4      7.230e+03  1.789e+03   4.040 6.31e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11780 on 434 degrees of freedom
Multiple R-squared:  0.9432,	Adjusted R-squared:  0.9427 
F-statistic:  1801 on 4 and 434 DF,  p-value: < 2.2e-16
```

```
> CDI2$fitted=model1$fitted
> ggplot(CDI2,aes(x=pop,y=fitted,color=region))+
+   geom_line()+
+   geom_point(aes(x=pop,y=crime,shape=region))
```



다중회귀분석

교호작용

범주형 변수와 연속형 변수 간의 교호작용

- 지역에 따라 인구 규모가 범죄건수에 미치는 영향의 정도가 다를까?

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_{2,2} + \beta_3 X_{2,3} + \beta_4 X_{2,4} + \beta_5 X_1 X_{2,2} + \beta_6 X_1 X_{2,3} + \beta_7 X_1 X_{2,4} + \epsilon$$

- $X_{2,2}$=1 if region=2, 0 if otherwise
- $X_{2,3}$=1 if region=3, 0 if otherwise
- $X_{2,4}$=1 if region=4, 0 if otherwise

- NE region: $E(y) = \beta_0 + \beta_1 X_1$
- NC region: $E(y) = (\beta_0 + \beta_2) + (\beta_1 + \beta_5) X_1$
- S region: $E(y) = (\beta_0 + \beta_3) + (\beta_1 + \beta_6) X_1$
- W region: $E(y) = (\beta_0 + \beta_4) + (\beta_1 + \beta_7) X_1$

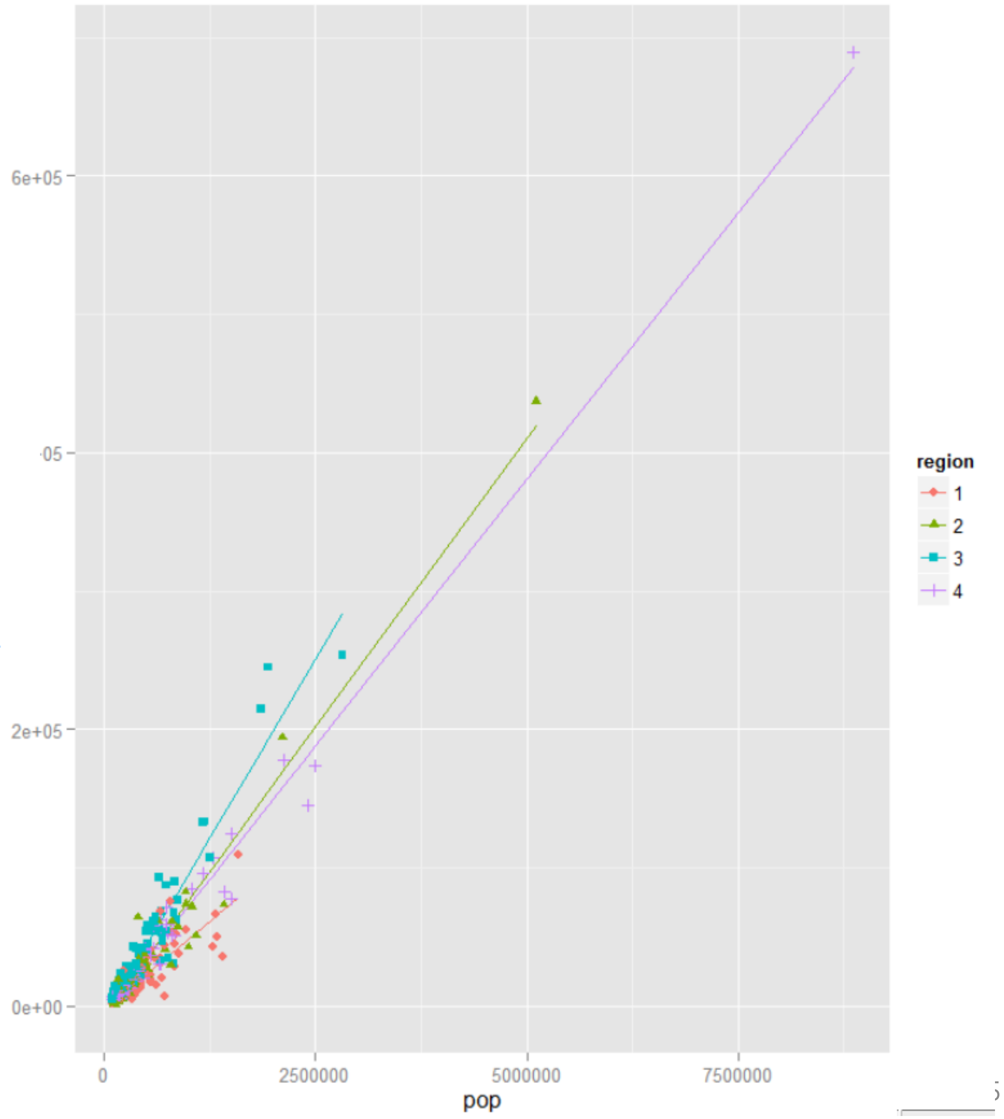
- β_5: pop의 기울기가 NE에 비해 NC 지역에서 얼마나 높은가
- β_6: pop의 기울기가 NE에 비해 S 지역에서 얼마나 높은가
- β_7: pop의 기울기가 NE에 비해 W 지역에서 얼마나 높은가

```
> model2=lm(crime~pop+region+pop*region,CDI2)
> summary(model2)

Call:
lm(formula = crime ~ pop + region + pop * region, data = CDI2)

Residuals:
   Min     1Q Median     3Q    Max
-47267  -2797    455   3245  51931

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.781e+03  1.489e+03  -1.869   0.0624 .
pop          5.148e-02  3.022e-03  17.035  < 2e-16 ***
region2     -4.378e+03  1.850e+03  -2.367   0.0184 *
region3     -4.150e+03  1.830e+03  -2.267   0.0239 *
region4     -1.778e+03  1.945e+03  -0.914   0.3612
pop:region2  3.212e-02  3.459e-03   9.285  < 2e-16 ***
pop:region3  5.162e-02  3.733e-03  13.830  < 2e-16 ***
pop:region4  2.554e-02  3.191e-03   8.003 1.14e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9669 on 431 degrees of freedom
Multiple R-squared:  0.962,	Adjusted R-squared:  0.9614
F-statistic:  1559 on 7 and 431 DF,  p-value: < 2.2e-16
```

- 지역별로 기울기의 차이가 있는가?

$$H_0 : \beta_5 = \beta_6 = \beta_7 = 0$$

```
> anova(model2)
Analysis of Variance Table

Response: crime
             Df     Sum Sq    Mean Sq   F value    Pr(>F)    
pop           1 9.8775e+11 9.8775e+11 10564.546 < 2.2e-16 ***
region        3 1.2430e+10 4.1433e+09    44.316 < 2.2e-16 ***
pop:region    3 1.9943e+10 6.6478e+09    71.102 < 2.2e-16 ***
Residuals   431 4.0297e+10 9.3496e+07                        
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



연속형 변수 간의 교호작용

- 인구 규모와 실업률이 범죄 건수에 미치는 영향은?
- 실업률이 높을 수록 인구 규모가 범죄 건수에 미치는 영향이 커질까?

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

$$y = \beta_0 + \beta_2 X_2 + (\beta_1 + \beta_3 X_2) X_1 + \epsilon$$

$$y = \beta_0 + \beta_1 X_1 + (\beta_2 + \beta_3 X_1) X_2 + \epsilon$$

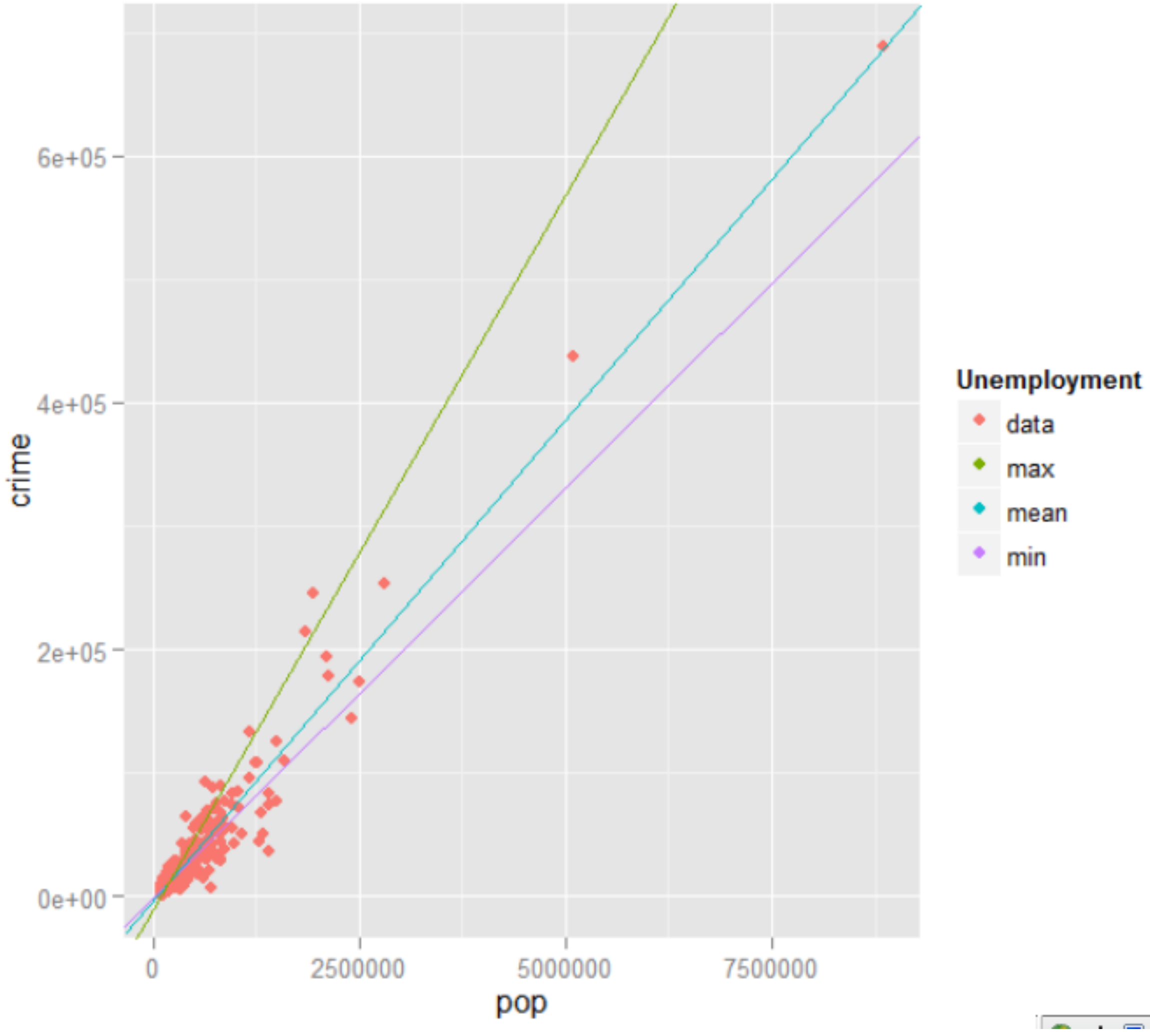
Crime=-170.8-449.2*unemployment +
(608.5+0.002598*unemployment)*pop

- 동일한 실업률 수준이 유지될 때 인구가 증가할 수록 범죄 건수가 증가함
- 그 기울기가 실업률이 높을 수록 가파름

주효과가 유의하지 않더라도 교호작용이 유의하면 제거하지 않음

```
> model3=lm(crime~pop+unemployment+pop*unemployment,CDI2)
> summary(model3)

Call:
lm(formula = crime ~ pop + unemployment + pop * unemployment,
    data = CDI2)

Residuals:
   Min     1Q Median     3Q    Max
-70804  -3685    -53   3683  88690

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.708e+03  2.556e+03  -0.668 0.504429
pop               6.085e-02  5.530e-03  11.002  < 2e-16 ***
unemployment     -4.492e+02  3.508e+02  -1.281 0.201040
pop:unemployment  2.598e-03  7.482e-04   3.472 0.000567 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12720 on 435 degrees of freedom
Multiple R-squared:  0.9336,	Adjusted R-squared:  0.9332
F-statistic:  2039 on 3 and 435 DF,  p-value: < 2.2e-16
```

```
ggplot(CDI2, aes(x=pop,y=crime,colour="data"))+
  geom_point()+
  geom_abline(intercept=-1.708e+03 + -4.492e+02 *2.2 , slope= 6.085e-02 +  2.598e-03 *2.2,aes(colour="min"))+
  geom_abline(intercept=-1.708e+03 + -4.492e+02 *6.59 , slope= 6.085e-02 +  2.598e-03 *6.59,aes(colour="mean"))+
  geom_abline(intercept=-1.708e+03 + -4.492e+02 *21.3 , slope= 6.085e-02 +  2.598e-03 *21.3,aes(colour="max"))+
  scale_color_discrete(name="Unemployment")
```



```r
ggplot(CDI2, aes(x=unemployment,y=crime,colour="data"))+
  geom_point()+
  geom_abline(intercept=-1.708e+03 + 6.085e-02 * 100000 , slope=-4.492e+02+  2.598e-03 *100000,aes(colour="min"))+
  geom_abline(intercept=-1.708e+03 +6.085e-02 * 388700 , slope=-4.492e+02 +  2.598e-03 *388700,aes(colour="mean"))+
  geom_abline(intercept=-1.708e+03 + 6.085e-02 *8863000 , slope= -4.492e+02 +  2.598e-03 *8863000,aes(colour="max"))+
  scale_color_discrete(name="Population")
```

Crime=-170.8+0.06085*pop+ (-449.2+0.002598*pop)*unemployment

- 동일한 인구 규모가 유지될 때 실업률이 증가할 수록 범죄 건수는 증가
- 그 기울기는 인구 규모가 증가할 수록 가파름