회귀모형의 적합 및 추론

1. 단순선형회귀모형 적합

- 반응변수 Y와 설명변수 X 사이의 선형관계 가정

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad i = 1, \ldots, n$$

  - 일차적인 관심: 회귀계수 β_0와 β_1의 추정
  - 오차항 ε_i에 대한 가정: 서로 독립, 동일 분포 $N(0, \sigma^2)$

- 예제: 데이터 프레임 women
  - 변수 height와 weight의 관계 탐색
  - 첫 번째 작업: 산점도 작성

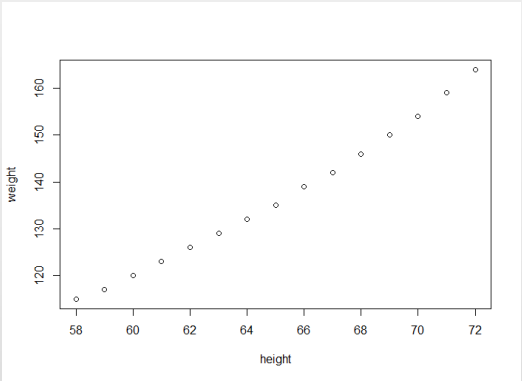
- 두 변수의 산점도 작성

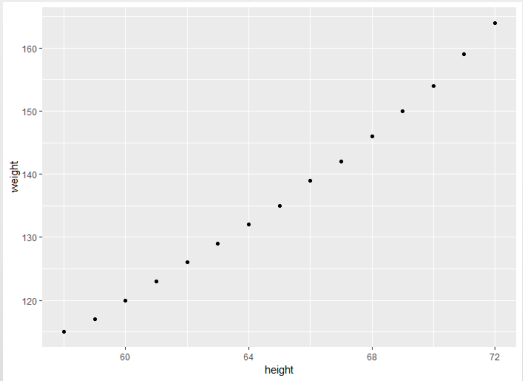
  - base graphics에 의한 산점도

```
> plot(weight ~ height, women)
```

  - ggplot2에 의한 산점도

```
> library(ggplot2)
> ggplot(women, aes(x=height, y=weight)) +
    geom_point()
```





geom 함수: 그래프 작성 함수

geom_point( )
geom_line( )
geom_smooth( )
geom_smooth(method="lm")

선형관계가 있는 것으로 보임

- 선형회귀모형 적합: 함수 lm()

```
> fit <- lm(weight ~ height, women)

> fit

Call:
lm(formula = weight ~ height, data = women)

Coefficients:
(Intercept)       height
     -87.52         3.45
```

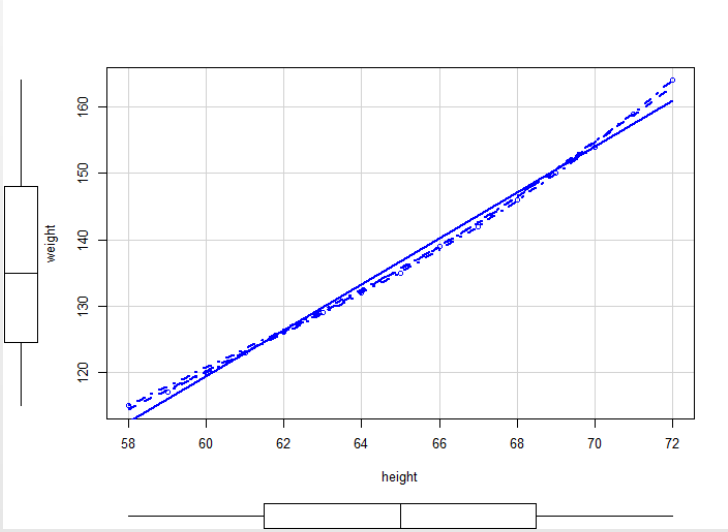
```
> names(fit)
 [1] "coefficients"  "residuals"   "effects"     "rank"
 [5] "fitted.values" "assign"      "qr"          "df.residual"
 [9] "xlevels"       "call"        "terms"       "model"
```

- 사용자마다 필요한 정보가 서로 다를 수 있음
- 필요한 정보를 각자 선택해서 추출
- 모든 결과를 한번에 출력하는 SAS, SPSS와는 다른 접근 방식

- 개선된 형태의 두 변수 산점도

  - 패키지 car의 함수 scatterplot()

```
> library(car)
> scatterplot(weight ~ height, women)
```



- 두 변수의 산점도
- 두 변수의 상자그림

- 회귀직선
- 비모수 회귀곡선

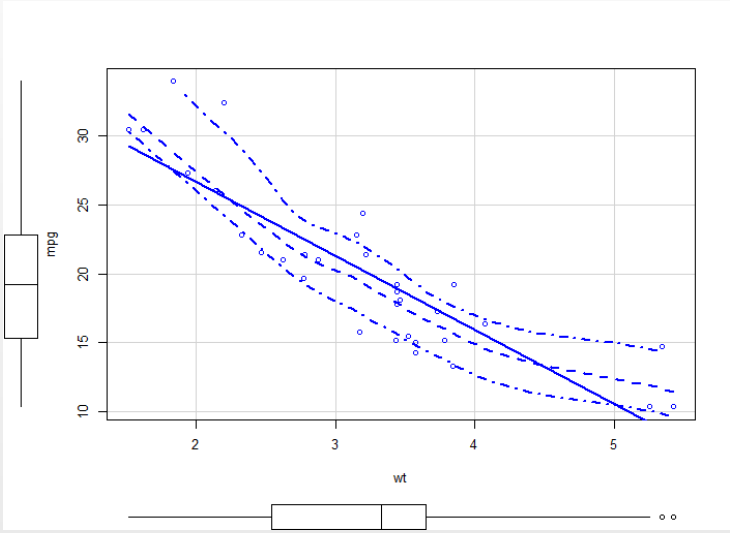
- 자료의 분산 추정과 관련된 두 개의 비모수 회귀곡선

- 분산 추정과 관련된 두 곡선이 겹쳐져 있어 구분이 조금 어려움

- 다른 예제로 형태 확인

- 데이터 프레임 mtcars에서 변수 mpg와 wt 적용

```
> scatterplot(mpg ~ wt, mtcars)
```



- 회귀직선: 실선(lty=1)
- 비모수 곡선: dashed line(lty=2)

- 분산 추정 관련 두 비모수 곡선: lty=4

2. 다중선형회귀모형 적합

- 반응변수 Y와 설명변수 $X_1, X_2, \ldots, X_k$ 사이에 선형 관계 가정

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \varepsilon_i, \qquad i = 1, \ldots, n$$

  - 오차항 ε_i 가정: 서로 독립, 같은 분포 $N(0, \sigma^2)$

- 선형 및 오차항 가정
  - 회귀모형의 추정 및 추론의 정당성 보장
  - 가정 위반 시 추론 결과가 부정확하게 나올 수 있음

- 함수 lm()의 기본적인 사용법

```
lm(formula, data, subset, weights, …)
```

  - formula: 회귀모형 설정을 위한 R 공식
  - data: 데이터 프레임
  - weights: 각 관찰값에 가중값 부여하는 경우
  - subset: 데이터의 일부분만을 이용하는 경우
    처음 100개 자료만 이용: `lm(y ~ x, subset=1:100)`
    변수 z가 0 이상인 자료만 이용: `lm(y ~ x, subet= z>=0)`

- R 공식에 사용되는 기호

1) 물결표(~): 반응변수 ~ 설명변수
2) 플러스(+): 설명변수 구분. `y ~ x1 + x2 + x3`
3) 콜론(:): 설명변수 사이의 상호작용. `y ~ x1 + x2 + x1:x2`
4) 별표(*): 모든 가능한 상호 작용. `y ~ x1*x2` → `y ~ x1 + x2 + x1:x2`
5) 마침표(.): 반응변수를 제외한 데이터 프레임에 있는 모든 변수. 데이터 프레임에 y, x1, x2, x3가 있다면 `y~.` → y ~ x1 + x2 + x3
6) 마이너스(-): 모형에서 제외되는 변수
7) `- 1` 또는 `+ 0`: 절편 제거
8) `I()`: 괄호 안의 연산자를 수학 연산자로 인식. y~I(x1+x2) → $Y = \beta_0 + \beta_1(X_1 + X_2)$
9) poly(x, n): 변수 x의 n차 다항회귀모형

- 예제 1: 행렬 state.x77
  - 미국 50개 주와 관련된 8개 변수로 구성된 행렬
  - 반응변수: Murder
  - 설명변수: Population, Illiteracy, Income, Frost

• 행렬을 데이터 프레임으로 전환하고, 필요한 변수만을 선택

```
> states <- as.data.frame(state.x77)
> states <- subset(states,
  select=c(Murder, Population, Illiteracy, Income, Frost))
```

```
> library(dplyr)
> states <- as.data.frame(state.x77)
> states <- select(states, Murder, Population,
                   Illiteracy, Income, Frost)
```

```
> head(states, n=3)
        Murder Population Illiteracy Income Frost
Alabama   15.1       3615        2.1   3624    20
Alaska    11.3        365        1.5   6315   152
Arizona    7.8       2212        1.8   4530    15
```

- 모형에 포함될 변수들의 관계 탐색
  - 상관계수
  - 산점도 행렬

- 상관계수 계산: 함수 cor()

```
cor(x, y=NULL, use="everything",
                    method=c("pearson", kendall", "spearman"))
```

  - x, y: 벡터, 행렬, 데이터 프레임
    x만 있는 경우: x에 있는 모든 변수들 사이의 상관계수 계산
    x와 y가 있는 경우: x에 있는 변수와 y에 있는 변수를 하나씩 짝을 지어 상관계수 계산
  - use: 결측값 처리 방식.
    "everything": 결측값이 있으면 NA
    "pairwise": 상관계수가 계산되는 변수만을 대상으로 NA가 있는 케이스 제거
  - method: 상관계수의 종류

- 데이터 프레임 states에 있는 변수들의 상관계수

```
> cor(states)
               Murder Population Illiteracy     Income       Frost
Murder      1.0000000  0.3436428  0.7029752 -0.2300776 -0.5388834
Population  0.3436428  1.0000000  0.1076224  0.2082276 -0.3321525
Illiteracy  0.7029752  0.1076224  1.0000000 -0.4370752 -0.6719470
Income     -0.2300776  0.2082276 -0.4370752  1.0000000  0.2262822
Frost      -0.5388834 -0.3321525 -0.6719470  0.2262822  1.0000000
```

  - 상관계수 행렬: 변수의 개수가 많아지면 변수 사이 관계 파악이 어려움
  - 상관계수 행렬을 그래프로 표현: 패키지 GGally의 함수 `ggcorr()`

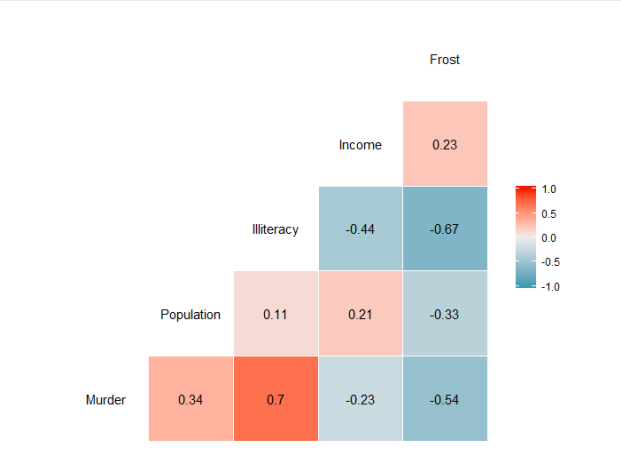
- 패키지 GGally의 함수 ggcorr()

```
ggcorr(data, method=c("pairwise", pearson"), label=FALSE,
                label_round=1, …)
```

  - label: 그래프에 상관계수 표시 여부
  - label_round: 상관계수 반올림 자릿수

- states 변수들의 상관계수 그래프

```
> library(GGally)
> ggcorr(states, label=TRUE, label_round=2)
```



- 산점도 행렬

  - 여러 변수로 이루어진 자료에서 두 변수끼리 짝을 지어 작성된 산점도를 행렬 형태로 배열

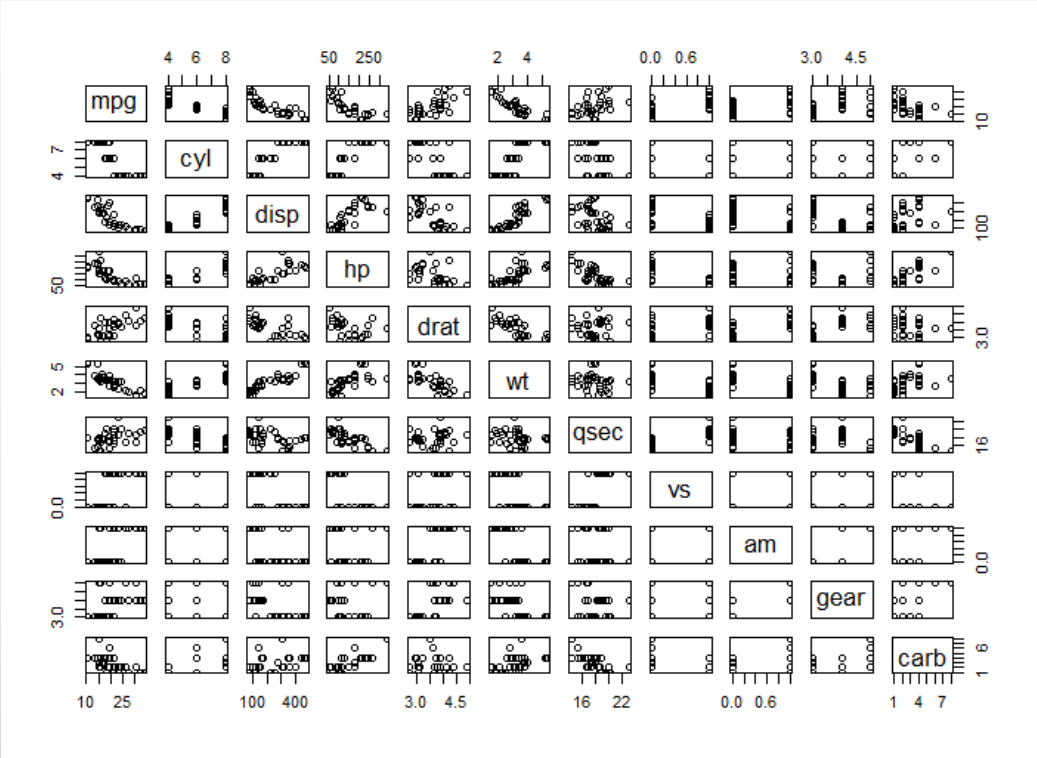
  - 회귀분석에서 필수적인 그래프

  - 작성 방법
    1) 패키지 graphics의 함수 `pairs()`
    2) 패키지 GGally의 함수 `ggpairs()`
    3) 패키지 car의 함수 `scatterplotMatrix()`

- 함수 pairs()

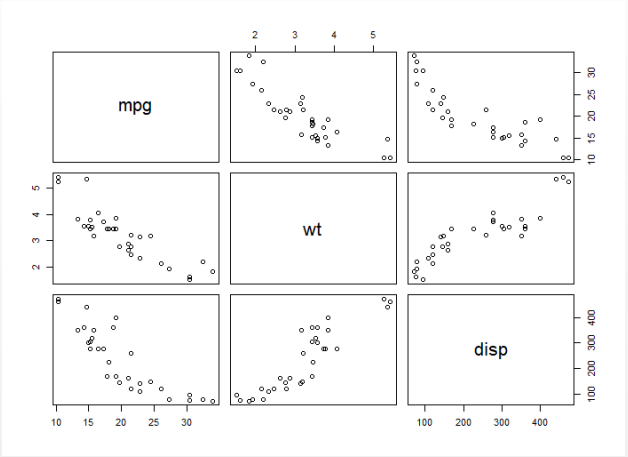
```
> pairs(mtcars)
```

- 입력된 데이터 프레임의 모든 변수에 대한 산점도 행렬 작성
- 변수가 많은 경우에는 의미 없는 그래프



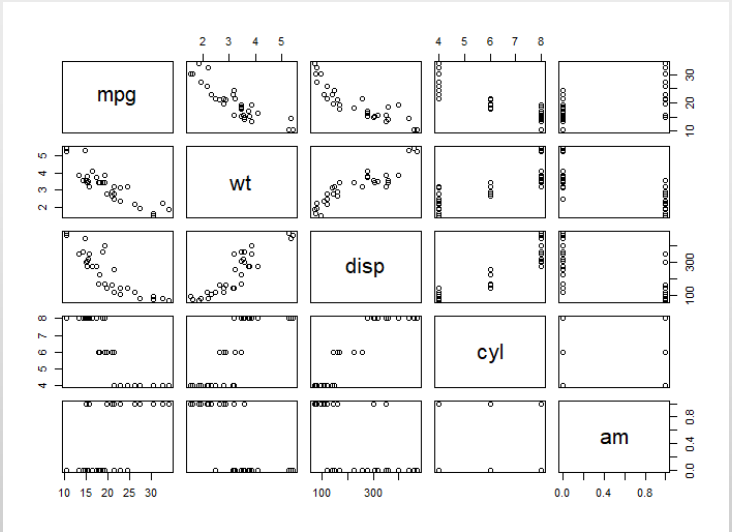
- 필요한 변수 선택하여 산점도 행렬 작성

```
> pairs(~mpg+wt+disp, data=mtcars)
```



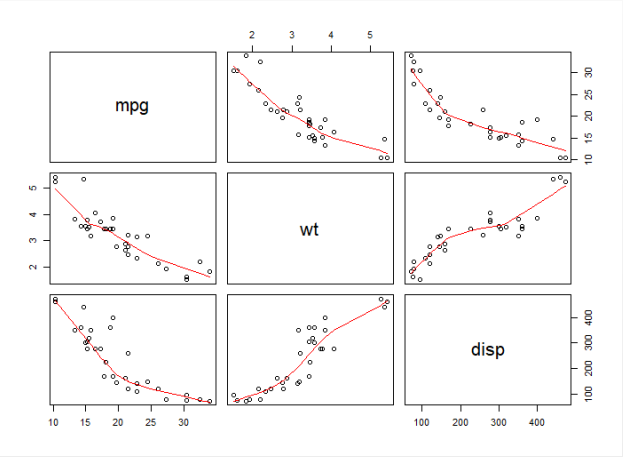

```
> library(dplyr)
> mtcars_1 <- select(mtcars, mpg, wt, disp, cyl, am)
> pairs(mtcars_1)
```

- mpg: 연비
- wt: 무게
- disp: 배기량
- cyl: 실린더 개수
- am: 변속기 종류



- 패널 함수 이용

```
> pairs(~mpg+wt+disp, data=mtcars, panel=panel.smooth)
```



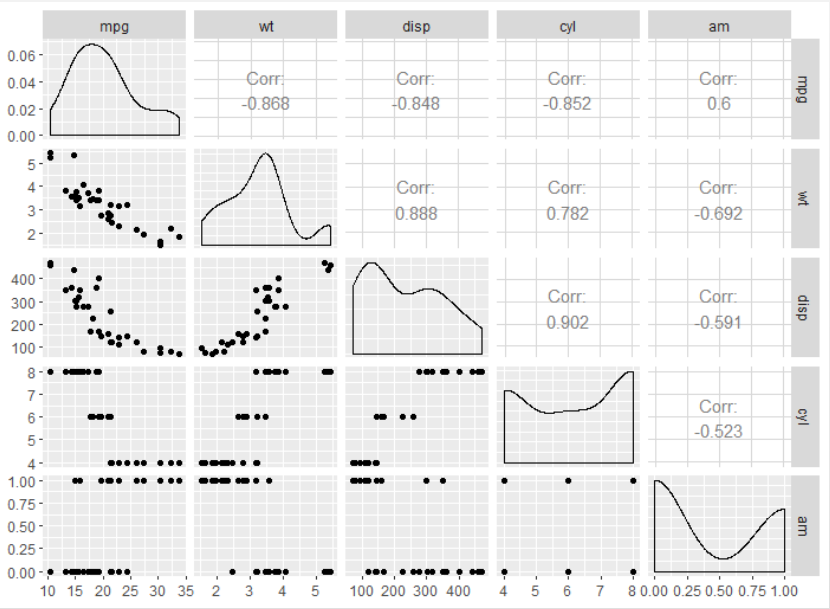
- panel에 패널 함수 지정
- panel.smooth: 로버스트 국소선형회귀 곡선을 그리는 패널 함수

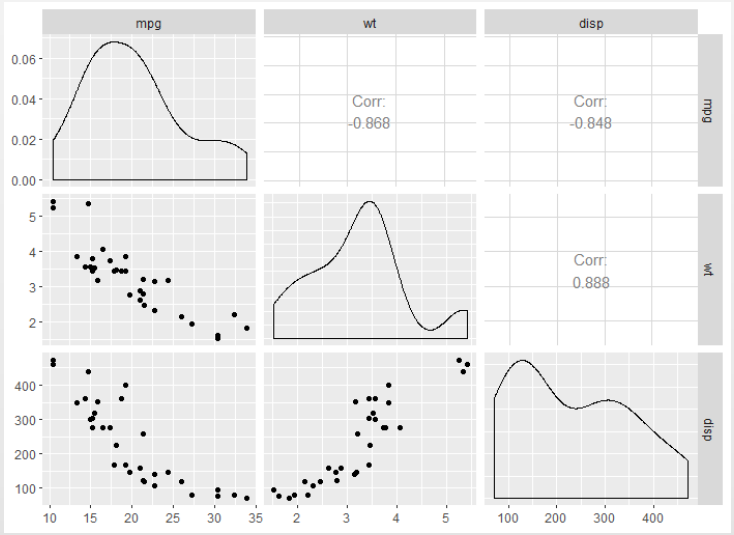
- 패키지 GGally의 함수 ggpairs()

```
> library(GGally)
> ggpairs(mtcars_1)
```

- 사용될 변수만으로 이루어진 데이터 프레임 입력

```
> ggpairs(mtcars_1, columns=1:3)
```

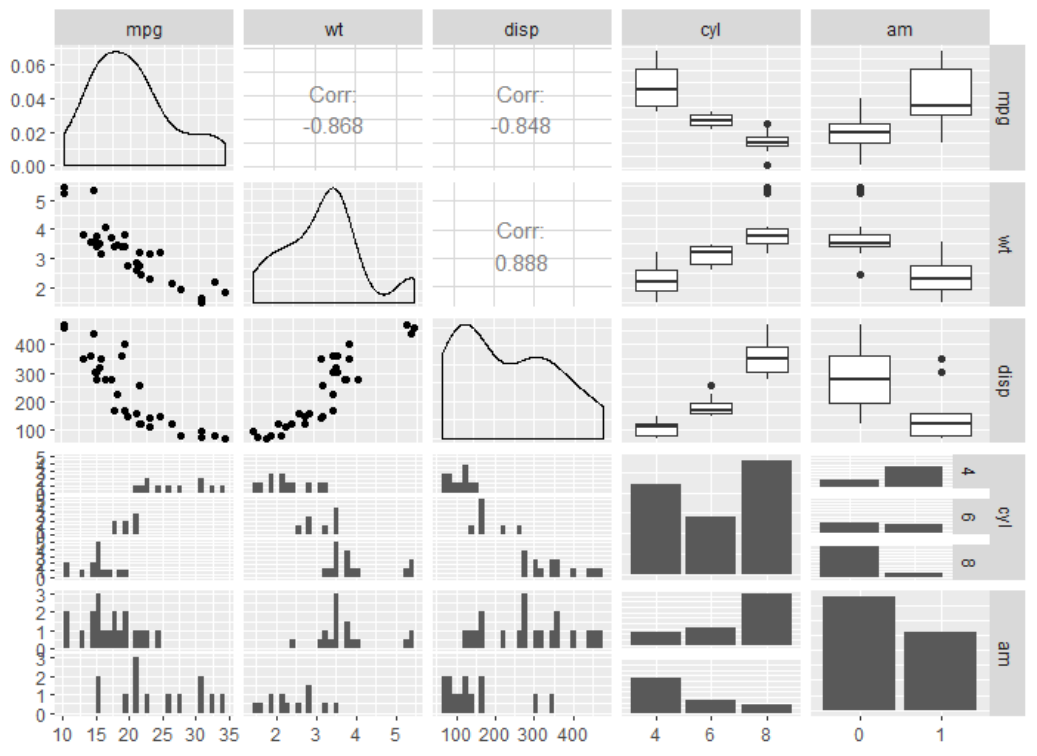




- 모든 변수가 숫자형
- 변수 cyl과 am을 요인으로 전환하고 다시 작성

- 숫자형과 요인이 함께 있는 경우

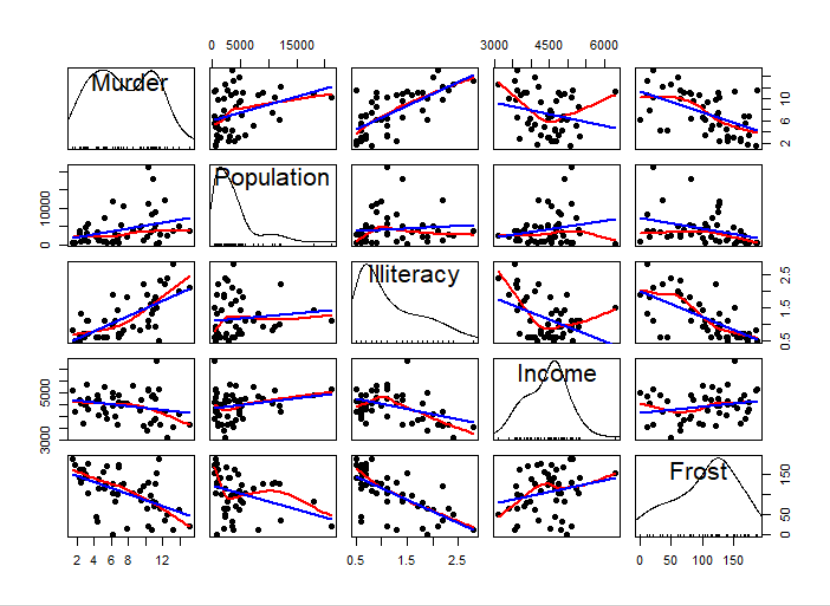
```
> library(dplyr)
> mtcars_2 <- mutate(mtcars_1, am=factor(am), cyl=factor(cyl))
> ggpairs(mtcars_2)
```



- 패키지 car의 함수 `scatterplotMatrix()`

  - 데이터 프레임 입력
  - 산점도의 색, 모양 조절 가능: col, pch
  - 패널에 나타나는 회귀직선 조절: regLine에 리스트로 지정
  - 패널에 나타나는 비모수 회귀곡선 조절: smooth에 리스트로 지정

```
> library(car)
> scatterplotMatrix(states, col="black", pch=19,
    regLine=list(lty=1, col="blue"),
    smooth=list(spread=FALSE, lty.smooth=1, col.smooth="red"))
```



- 예제 1 계속: states에 대한 회귀모형 적합

```
> fit <- lm(Murder~Population+Illiteracy+Income+Frost, states)
> fit

Call:
lm(formula = Murder ~ Population + Illiteracy + Income + Frost,
    data = states)

Coefficients:
(Intercept)   Population   Illiteracy       Income        Frost
  1.235e+00    2.237e-04    4.143e+00    6.442e-05    5.813e-04
```

- 함수 lm()으로 생성된 객체(회귀분석 결과)의 내용 확인을 위한 함수
  - anova(): 분산분석표
  - coefficients(): 추정된 회귀계수, coef()도 가능
  - confint(): 회귀계수 신뢰구간.
  - fitted(): 반응변수 적합값
  - residuals(): 잔차. resid()도 가능
  - summary(): 중요한 적합 결과 요약

- 예제 2: women

  - 데이터 프레임 women의 변수 weight와 height의 관계
  - 선형보다는 2차가 더 적합한 것으로 보임

  - 다항회귀모형

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_p X_i^p + \varepsilon_i$$

  - 차수 p를 너무 높이면 다중공선성의 문제가 발생할 수 있음
  - 3차를 넘지 않는 것이 일반적

  - R 함수: poly(x, degree=1, raw=FALSE)
    degree: 차수 지정
    raw: 직교다항회귀 여부.
    일반적인 다항 회귀의 경우는 TRUE

- 반응변수 weight에 대한 height의 2차 다항회귀모형 적합

```
> fit_w <- lm(weight ~ poly(height, degree=2, raw=TRUE), women)
> fit_w

Call:
lm(formula = weight ~ poly(height, degree = 2, raw = TRUE),
data = women)

Coefficients:
                            (Intercept)
                              261.87818
poly(height, degree = 2, raw = TRUE)1
                               -7.34832
poly(height, degree = 2, raw = TRUE)2
                                0.08306
```

```
> fit_w <- lm(weight ~ height + I(height^2), women)
```

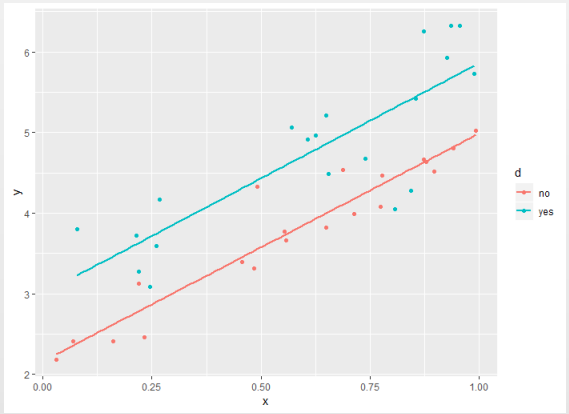
모형식: $\hat{Y}_i = 261.87 - 7.345X_i + 0.083X_i^2$

- 예제 3: 질적 변수를 설명변수로 사용

  - 회귀모형에서 사용되는 변수 형태
    반응변수: 연속형(정규분포 가정 필요)
    설명변수: 연속형(정규분포 가정은 필요 없으나, 가능한 좌우대칭)
    범주형(가변수 필요)

  - 가변수 회귀모형: 2개 범주(yes, no) → 1개 가변수

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \varepsilon, \quad D = \begin{cases} 0 & no \\ 1 & yes \end{cases}$$



  - D=0인 범주: 기준 범주
  - 회귀계수 β_2: yes 범주와 기준 범주의 차이

  - 일반적으로 가변수 개수=범주 개수-1

  - 만일 가변수 개수=범주 개수이면 회귀계수 추정이 불가능
    → 절편 제거하면 추정 가능
    → 회귀계수의 해석이 달라짐(해당 범주의 효과)
    → 두 개 이상의 범주형 변수가 포함되는 경우에는 적용이 어려움

- 패키지 carData의 데이터 프레임 Leinhardt
  - 1970년대 105개 나라의 신생아 사망률, 소득, 지역 및 원유 수출 여부
  - 반응변수: 신생아 사망률(infant)
  - 설명변수: 소득(income), 지역(region, 4개 수준: Africa, Americas, Asia, Europe), 원유 수출(oil, 2개 수준: no, yes)

- 함수 lm()에 요인 입력: 자동으로 필요한 개수의 가변수 포함

```
> lm(infant ~ income + region, data=Leinhardt)

Call:
lm(formula = infant ~ income + region, data = Leinhardt)

Coefficients:
   (Intercept)          income  regionAmericas       regionAsia
     1.432e+02      -3.458e-03      -8.473e+01      -4.480e+01
  regionEurope
    -1.135e+02
```

  - 기준 범주: 알파벳 첫 번째 범주인 Africa
  - 회귀계수 regionAmericas는 범주 Americas와 기준 범주 Africa의 차이

- 절편 제거 모형: + 0 또는 -1 포함

```
> lm(infant ~ income + region + 0, data=Leinhardt)

Call:
lm(formula = infant ~ income + region + 0, data = Leinhardt)

Coefficients:
        income    regionAfrica  regionAmericas      regionAsia
     -0.003458      143.235952       58.504549       98.440309
  regionEurope
     29.767837
```

- 두 범주형 변수(region, oil) 포함

```
> lm(infant ~ income + region + oil, data=Leinhardt)

Call:
lm(formula = infant ~ income + region + oil, data = Leinhardt)

Coefficients:
   (Intercept)          income  regionAmericas      regionAsia
     136.82468        -0.00529       -83.64943       -45.88540
  regionEurope          oilyes
    -101.48624        78.33508
```

3. 회귀모형의 추론

- 회귀모형: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$

- 회귀계수에 대한 가설
  1) $H_0: \beta_1 = \cdots = \beta_k = 0$
  2) $H_0: \beta_q = \beta_{q+1} = \cdots = \beta_r = 0, \quad q < r \leq k$
  3) $H_0: \beta_i = 0, \ H_1: \beta_i \neq 0$

- 회귀계수의 신뢰구간

- 회귀모형 적합 정도에 대한 통계량: 결정계수, 수정된 결정계수, MSE, …

- 예측
  1) 반응변수의 평균값에 대한 예측
  2) 반응변수의 개별 관찰값에 대한 예측

- 적합된 회귀모형 추론을 위한 함수

  - 함수 summary()

```
> fit1 <- lm(Murder~Population+Illiteracy+Income+Frost, states)
> summary(fit1)
```

```
Call:
lm(formula = Murder ~ Population + Illiteracy + Income + Frost,
    data = states)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7960 -1.6495 -0.0811  1.4815  7.6210

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.235e+00  3.866e+00   0.319   0.7510
Population  2.237e-04  9.052e-05   2.471   0.0173 *
Illiteracy  4.143e+00  8.744e-01   4.738 2.19e-05 ***
Income      6.442e-05  6.837e-04   0.094   0.9253
Frost       5.813e-04  1.005e-02   0.058   0.9541
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.535 on 45 degrees of freedom
Multiple R-squared:  0.567,     Adjusted R-squared:  0.5285
F-statistic: 14.73 on 4 and 45 DF,  p-value: 9.133e-08
```

- 개별 회귀계수 추정 및 검정
- $\sqrt{MSE}$
- 결정계수 및 수정된 결정계수
- 모든 회귀계수의 유의성 검정

- 함수 anova()

  - 적합된 회귀모형의 분산분석표

```
> anova(fit1)
Analysis of Variance Table

Response: Murder
            Df  Sum Sq Mean Sq F value   Pr(>F)
Population   1  78.854  78.854 12.2713 0.001052 **
Illiteracy   1 299.646 299.646 46.6307 1.83e-08 ***
Income       1   0.057   0.057  0.0089 0.925368
Frost        1   0.021   0.021  0.0033 0.954148
Residuals   45 289.167   6.426
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- 두 회귀모형의 비교

1) 확장모형(Ω): $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$

2) 축소모형(ω): 다음의 귀무가설이 사실인 모형

$$H_0: \beta_q = \beta_{q+1} = \cdots = \beta_r = 0, \qquad q < r \leq k$$

RSS_Ω: 확장모형의 잔차제곱합
RSS_ω: 축소모형의 잔차제곱합

- 만일 $RSS_\omega - RSS_\Omega$가 적다면, 축소모형이 확장모형만큼 좋다는 의미
- 모수절약의 원칙에 따라 축소모형 선택 가능

- 검정통계량

$$F = \frac{(RSS_\omega - RSS_\Omega)/\text{두 모형의 모수 차이}}{RSS_\Omega/n-k-1}$$

- 함수 anova()
  - 두 회귀모형의 비교

```
> fit1 <- lm(Murder ~ ., states)
> fit2 <- lm(Murder ~ Population + Illiteracy, states)
```

```
> anova(fit2,fit1)
Analysis of Variance Table

Model 1: Murder ~ Population + Illiteracy
Model 2: Murder ~ Population + Illiteracy + Income + Frost
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     47 289.25
2     45 289.17  2  0.078505 0.0061 0.9939
```

anova(축소모형, 확장모형)

귀무가설의 기각이 어려움

- 함수 confint()

  - 모형에 포함된 회귀계수의 신뢰구간 추정
  - 사용법: confint(object, level=0.95)

95% 신뢰구간

```
> confint(fit1)
                    2.5 %       97.5 %
(Intercept) -6.552191e+00 9.0213182149
Population   4.136397e-05 0.0004059867
Illiteracy   2.381799e+00 5.9038743192
Income      -1.312611e-03 0.0014414600
Frost       -1.966781e-02 0.0208304170
```

90% 신뢰구간

```
> confint(fit1,level=0.9)
                      5 %         95 %
(Intercept) -5.258296e+00 7.7274224120
Population   7.165796e-05 0.0003756927
Illiteracy   2.674424e+00 5.6112492933
Income      -1.083794e-03 0.0012126432
Frost       -1.630309e-02 0.0174656982
```

- 예측

  - 회귀모형 적합 $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k$

  - 새롭게 주어진 설명변수 자료 $X_{1o}, X_{2o}, \ldots, X_{ko}$

  - 새롭게 주어진 설명변수 자료에 대한 반응변수의 값 예측

  - 반응변수의 예측
    1) 반응변수의 평균값에 대한 예측: $E(Y|X)$
    2) 반응변수의 개별 관찰값에 대한 예측: $Y_i = E(Y|X) + \varepsilon_i$

  - 두 경우의 예측에서
    예측 결과는 동일: $\hat{Y}_o = \hat{\beta}_0 + \hat{\beta}_1 X_{1o} + \cdots + \hat{\beta}_k X_{ko}$
    예측 오차는 다름: 개별 관찰값에 대한 경우가 더 크게 됨

- 함수 predict()

  - 반응변수의 예측
  - 사용법:

```
predict(object, newdata,
        interval=c("confidence", "prediction"), level=0.95)
```

object: lm object
newdata: 새롭게 주어진 설명변수 자료. 데이터 프레임
interval: 반응변수 평균 예측(confidence),
개별 관찰값 예측(prediction)
level: 예측수준

- 모형 fit2에 대하여 두 변수 (Population, Illiteracy)에 대한 새로운 관찰값 (15000, 0.8), (10000, 1.5), (5000, 2.5)
- 반응변수 Murder의 평균값 및 개별 관찰값 예측

새로운 자료 준비

```
> x0 <- data.frame(Population=c(15000,10000,5000),
                   Illiteracy=c(0.8,1.5,2.5))
```

평균 예측

```
> predict(fit2, newdata=x0, interval="confidence")
        fit       lwr      upr
1  8.278931  6.321113 10.23675
2 10.014516  8.820475 11.20856
3 12.974322 11.265395 14.68325
```

개별 관찰값 예측

```
> predict(fit2, newdata=x0, interval="prediction")
        fit      lwr      upr
1  8.278931 2.918002 13.63986
2 10.014516 4.883020 15.14601
3 12.974322 7.699197 18.24945
```