Regression

임요한

02/08/2024

Contents

0. 필요한 package 설치

```
# packages needed for this class
name_pkg <- c(
  "Hmisc", "psych", # For describe functions
  "ggplot2", # For ggplot function
  "GGally", #ggpairs, ggduo function
  "MASS", # For Boston data set
  "ISLR", # For Carseats data set
  "effects", # For effect function
  "dplyr", # For select function
  "olsrr", # For variable selection
  "knitr"
  )
bool_nopkg <- !name_pkg %in% rownames(installed.packages())
if (any(bool_nopkg)) {
  install.packages(name_pkg[bool_nopkg], repos = "http://cran.us.r-project.org")
}
# load multiple packages
invisible(lapply(name_pkg, library, character.only = T))
```

```
set.seed(1)
```

1. 광고 자료

자료 읽기

광고자료를 통해 회귀분석을 진행한다.

이를 위하여 광고자료를 불러들인다.

```
adv = read.csv("Advertising.csv", header=T, sep=",")
adv = adv[,-1]
names(adv) = tolower(names(adv))
str(adv)
```

```
## 'data.frame':    200 obs. of  4 variables:
##  $ tv       : num  230.1 44.5 17.2 151.5 180.8 ...
##  $ radio    : num  37.8 39.3 45.9 41.3 10.8 48.9 32.8 19.6 2.1 2.6 ...
##  $ newspaper: num  69.2 45.1 69.3 58.5 58.4 75 23.5 11.6 1 21.2 ...
##  $ sales    : num  22.1 10.4 9.3 18.5 12.9 7.2 11.8 13.2 4.8 10.6 ...
```

```
head(adv)
```

```
##      tv radio newspaper sales
## 1 230.1  37.8      69.2  22.1
## 2  44.5  39.3      45.1  10.4
## 3  17.2  45.9      69.3   9.3
## 4 151.5  41.3      58.5  18.5
## 5 180.8  10.8      58.4  12.9
## 6   8.7  48.9      75.0   7.2
```

1.1 자료 탐색 [무엇을 탐색하지?]

광고자료의 분석에 앞서 자료에 대하여 기본적인 파악이 필요하다.

“선형성”, “등분산성”, 정규성, 독립성

이를 탐색적 자료분석(EDA)라 부르며, 다음의 코드들을 통해 탐색적 자료분석을 진행한다.

```
summary(adv)
```

```
##        tv              radio          newspaper          sales
##  Min.   :  0.70   Min.   : 0.000   Min.   :  0.30   Min.   : 1.60
##  1st Qu.: 74.38   1st Qu.: 9.975   1st Qu.: 12.75   1st Qu.:10.38
##  Median :149.75   Median :22.900   Median : 25.75   Median :12.90
##  Mean   :147.04   Mean   :23.264   Mean   : 30.55   Mean   :14.02
##  3rd Qu.:218.82   3rd Qu.:36.525   3rd Qu.: 45.10   3rd Qu.:17.40
##  Max.   :296.40   Max.   :49.600   Max.   :114.00   Max.   :27.00
```

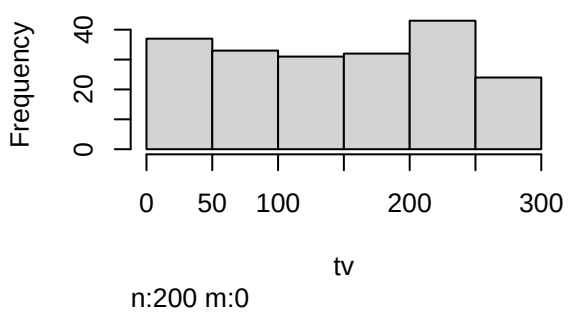
```
Hmisc::describe(adv)
```

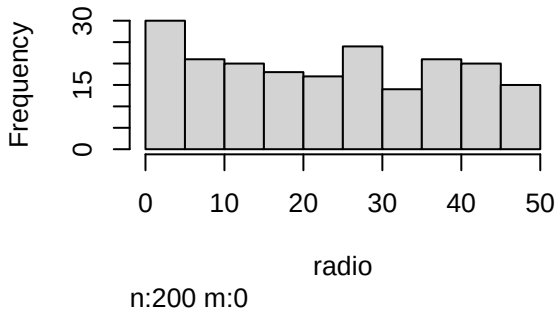
```
## adv
##
##  4  Variables      200  Observations
## ---------------------------------------------------------------------------
## tv
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      200        0      190        1      147    99.19    13.20    24.88
##      .25      .50      .75      .90      .95
##    74.38   149.75   218.82   261.44   280.74
##
## lowest :   0.7   4.1   5.4   7.3   7.8, highest: 289.7 290.7 292.9 293.6 296.4
## ---------------------------------------------------------------------------
## radio
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      200        0      167        1    23.26    17.16    1.995    3.400
##      .25      .50      .75      .90      .95
##    9.975   22.900   36.525   43.520   46.810
##
## lowest :  0.0  0.3  0.4  0.8  1.3, highest: 47.8 48.9 49.0 49.4 49.6
## ---------------------------------------------------------------------------
## newspaper
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      200        0      172        1    30.55    24.14     3.60     5.99
##      .25      .50      .75      .90      .95
##    12.75    25.75    45.10    59.07    71.82
##
## lowest :   0.3   0.9   1.0   1.7   1.8, highest:  79.2  84.8  89.4 100.9 114.0
## ---------------------------------------------------------------------------
## sales
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      200        0      121        1    14.02    5.907     6.60     7.96
##      .25      .50      .75      .90      .95
##    10.38    12.90    17.40    21.71    23.80
##
## lowest :  1.6  3.2  4.8  5.3  5.5, highest: 24.7 25.4 25.5 26.2 27.0
## ---------------------------------------------------------------------------
```

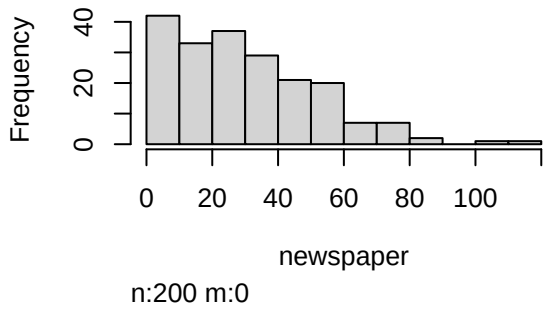
```
psych::describe(adv)
```

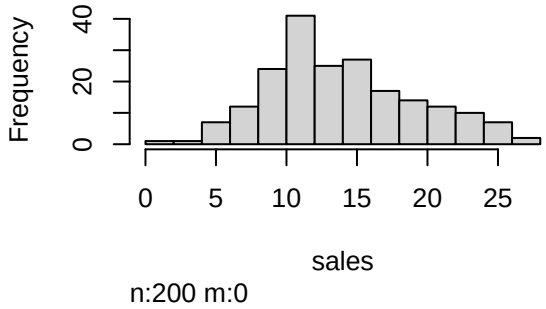
```
##           vars   n   mean    sd median trimmed    mad min   max range  skew
## tv           1 200 147.04 85.85 149.75  147.20 108.82 0.7 296.4 295.7 -0.07
## radio        2 200  23.26 14.85  22.90   23.00  19.79 0.0  49.6  49.6  0.09
## newspaper    3 200  30.55 21.78  25.75   28.41  23.13 0.3 114.0 113.7  0.88
## sales        4 200  14.02  5.22  12.90   13.78   4.82 1.6  27.0  25.4  0.40
##           kurtosis   se
## tv           -1.24 6.07
## radio        -1.28 1.05
## newspaper     0.57 1.54
## sales        -0.45 0.37
```

```
hist(adv)
```

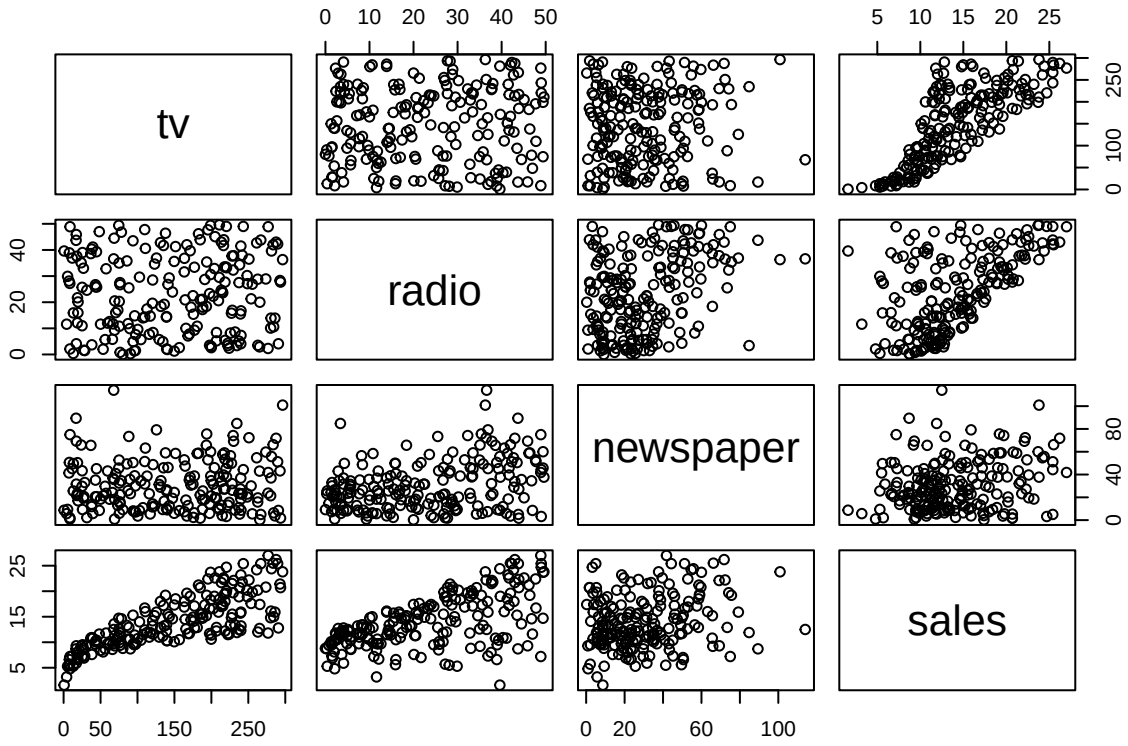








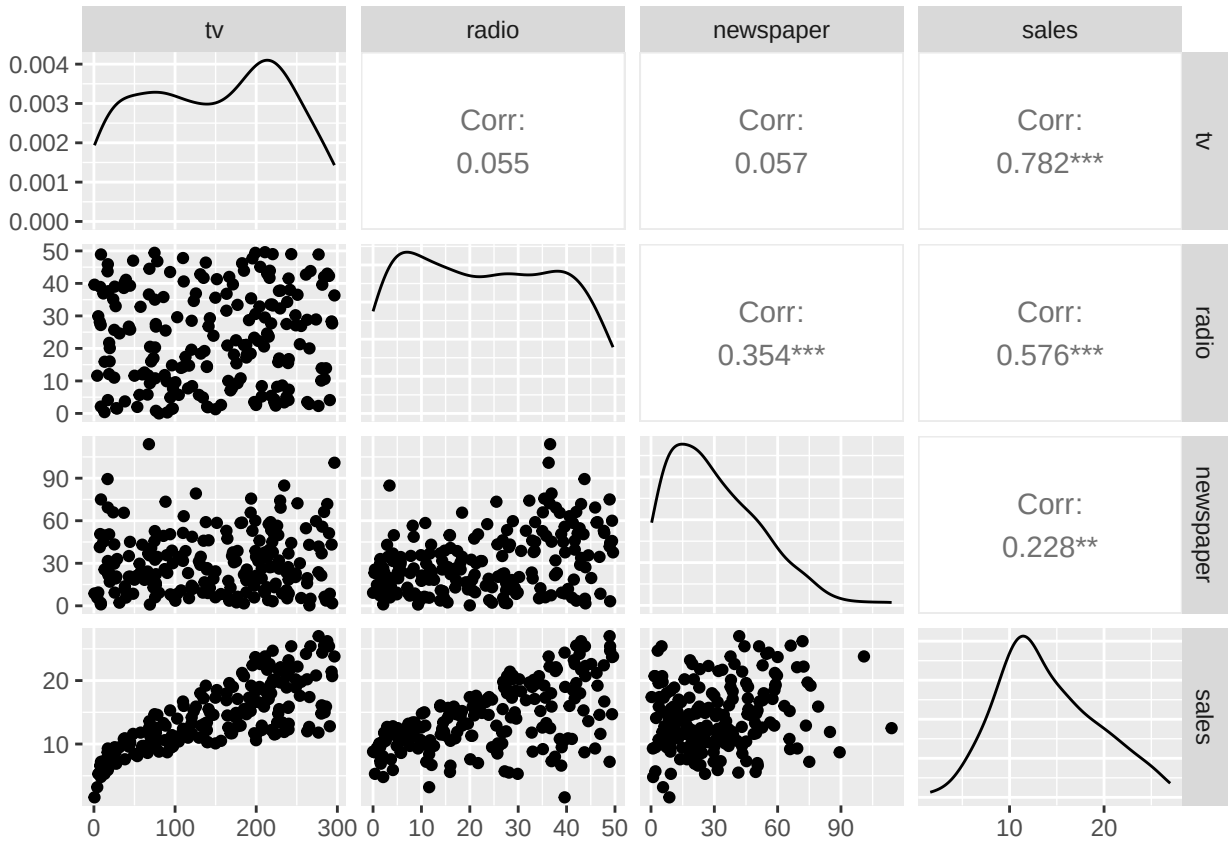

```
pairs(adv)
```



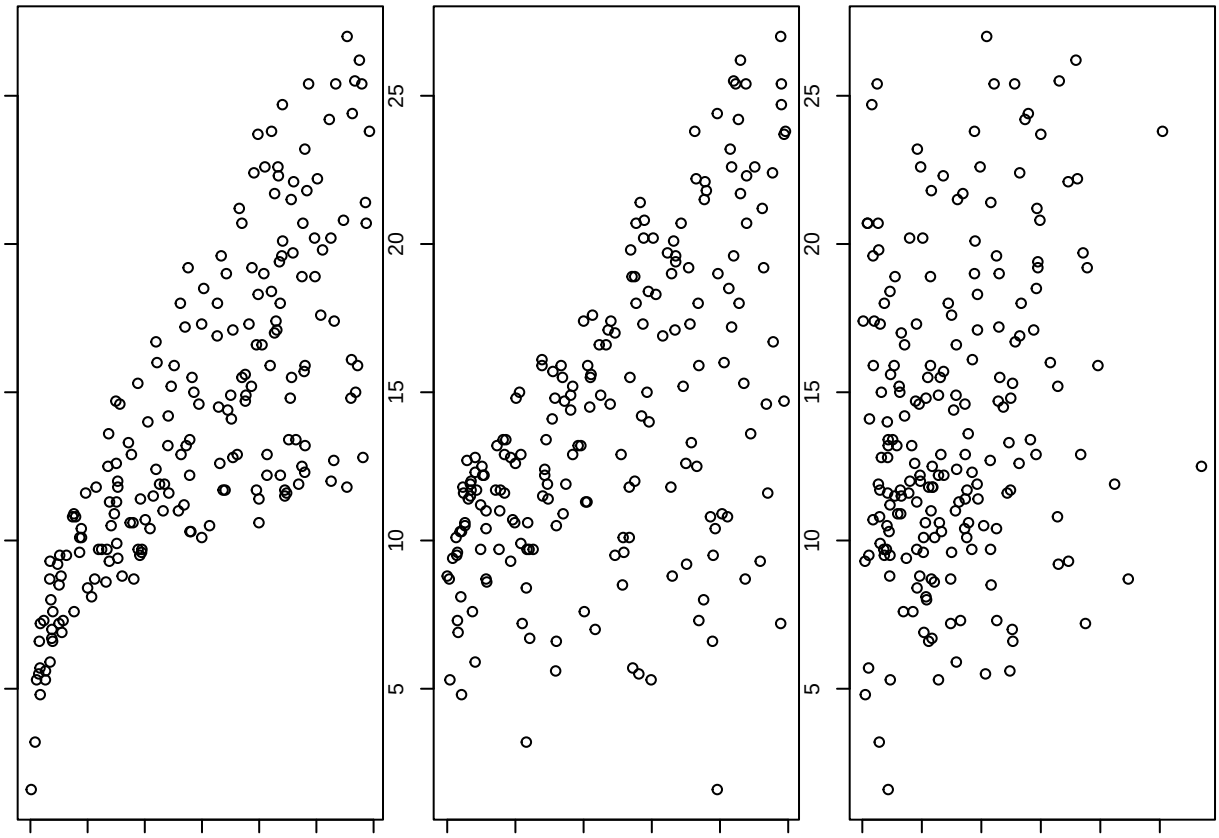

```
cor(adv)
```

```
##                    tv      radio  newspaper     sales
## tv         1.00000000 0.05480866 0.05664787 0.7822244
## radio      0.05480866 1.00000000 0.35410375 0.5762226
## newspaper 0.05664787 0.35410375 1.00000000 0.2282990
## sales      0.78222442 0.57622257 0.22829903 1.0000000
```

```
ggpairs(adv)
```




```
#ggplot(adv, aes(x=sales)) + geom_histogram(bins=10)

attach(adv)
#par("mar")
par(mar=c(1,1,1,1))
par(mfrow=c(1,3))
plot(tv,sales)
plot(radio,sales)
plot(newspaper,sales)
```




```
par(mfrow=c(1,1))
detach(adv)
```

1.2 단순 선형회귀모형의 적합

광고자료에서 반응변수가 sales일 경우의 단순선형 회귀분석을 진행한다.

이 때 독립변수는 각각 tv와 radio이다.

```
lm.fit = lm(sales ~ tv, data=adv)
str(lm.fit)
```

```
## List of 12
##  $ coefficients : Named num [1:2] 7.0326 0.0475
##   ..- attr(*, "names")= chr [1:2] "(Intercept)" "tv"
##  $ residuals    : Named num [1:200] 4.13 1.25 1.45 4.27 -2.73 ...
##   ..- attr(*, "names")= chr [1:200] "1" "2" "3" "4" ...
##  $ effects      : Named num [1:200] -198.31 57.57 1.08 3.99 -2.98 ...
##   ..- attr(*, "names")= chr [1:200] "(Intercept)" "tv" "" "" ...
##  $ rank         : int 2
##  $ fitted.values: Named num [1:200] 17.97 9.15 7.85 14.23 15.63 ...
##   ..- attr(*, "names")= chr [1:200] "1" "2" "3" "4" ...
##  $ assign       : int [1:2] 0 1
##  $ qr           :List of 5
##   ..$ qr   : num [1:200, 1:2] -14.1421 0.0707 0.0707 0.0707 0.0707 ...
##   .. ..- attr(*, "dimnames")=List of 2
##   .. .. ..$ : chr [1:200] "1" "2" "3" "4" ...
##   .. .. ..$ : chr [1:2] "(Intercept)" "tv"
##   .. ..- attr(*, "assign")= int [1:2] 0 1
##   ..$ qraux: num [1:2] 1.07 1.09
##   ..$ pivot: int [1:2] 1 2
##   ..$ tol  : num 1e-07
##   ..$ rank : int 2
##   ..- attr(*, "class")= chr "qr"
##  $ df.residual  : int 198
##  $ xlevels      : Named list()
##  $ call         : language lm(formula = sales ~ tv, data = adv)
##  $ terms        :Classes 'terms', 'formula'  language sales ~ tv
##   .. ..- attr(*, "variables")= language list(sales, tv)
##   .. ..- attr(*, "factors")= int [1:2, 1] 0 1
##   .. .. ..- attr(*, "dimnames")=List of 2
##   .. .. .. ..$ : chr [1:2] "sales" "tv"
##   .. .. .. ..$ : chr "tv"
##   .. ..- attr(*, "term.labels")= chr "tv"
##   .. ..- attr(*, "order")= int 1
##   .. ..- attr(*, "intercept")= int 1
##   .. ..- attr(*, "response")= int 1
##   .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
##   .. ..- attr(*, "predvars")= language list(sales, tv)
##   .. ..- attr(*, "dataClasses")= Named chr [1:2] "numeric" "numeric"
##   .. .. ..- attr(*, "names")= chr [1:2] "sales" "tv"
##  $ model        :'data.frame':   200 obs. of  2 variables:
##   ..$ sales: num [1:200] 22.1 10.4 9.3 18.5 12.9 7.2 11.8 13.2 4.8 10.6 ...
##   ..$ tv   : num [1:200] 230.1 44.5 17.2 151.5 180.8 ...
##   ..- attr(*, "terms")=Classes 'terms', 'formula'  language sales ~ tv
##   .. .. ..- attr(*, "variables")= language list(sales, tv)
##   .. .. ..- attr(*, "factors")= int [1:2, 1] 0 1
##   .. .. .. ..- attr(*, "dimnames")=List of 2
##   .. .. .. .. ..$ : chr [1:2] "sales" "tv"
```

```
##   .. .. .. .. ..$ : chr "tv"
##   .. .. ..- attr(*, "term.labels")= chr "tv"
##   .. .. ..- attr(*, "order")= int 1
##   .. .. ..- attr(*, "intercept")= int 1
##   .. .. ..- attr(*, "response")= int 1
##   .. .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
##   .. .. ..- attr(*, "predvars")= language list(sales, tv)
##   .. .. ..- attr(*, "dataClasses")= Named chr [1:2] "numeric" "numeric"
##   .. .. .. ..- attr(*, "names")= chr [1:2] "sales" "tv"
##  - attr(*, "class")= chr "lm"
```

```
summary(lm.fit)
```

```
##
## Call:
## lm(formula = sales ~ tv, data = adv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.032594   0.457843   15.36   <2e-16 ***
## tv          0.047537   0.002691   17.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

```
coef(lm.fit)
```

```
## (Intercept)          tv
##  7.03259355  0.04753664
```

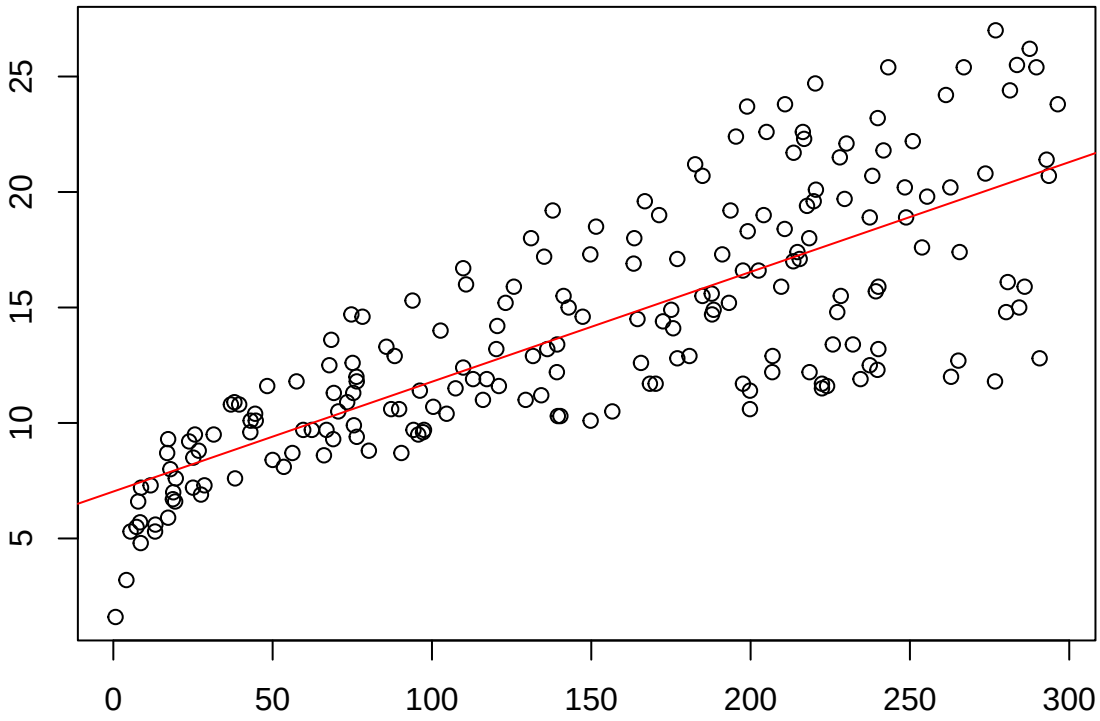
```
lm.fit2 = lm(sales ~ radio, data=adv)
summary(lm.fit2)
```

```
##
## Call:
## lm(formula = sales ~ radio, data = adv)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7305  -2.1324   0.7707   2.7775   8.1810
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.31164    0.56290  16.542   <2e-16 ***
## radio        0.20250    0.02041   9.921   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.275 on 198 degrees of freedom
```

```
## Multiple R-squared:  0.332,  Adjusted R-squared:  0.3287
## F-statistic: 98.42 on 1 and 198 DF,  p-value: < 2.2e-16
```

1.3 단순선형회귀 추정량의 특징

```
attach(adv)
par(mar=c(3,3,3,3))
plot(tv,sales)
abline(lm.fit,col="red")
```




```
detach(adv)
```

(1) 추정된 회귀선은 항상 (x-bar, y-bar)를 지난다.

(2) 최소제곱추정의 오차는 y축 방향의 오차이다. 왜?

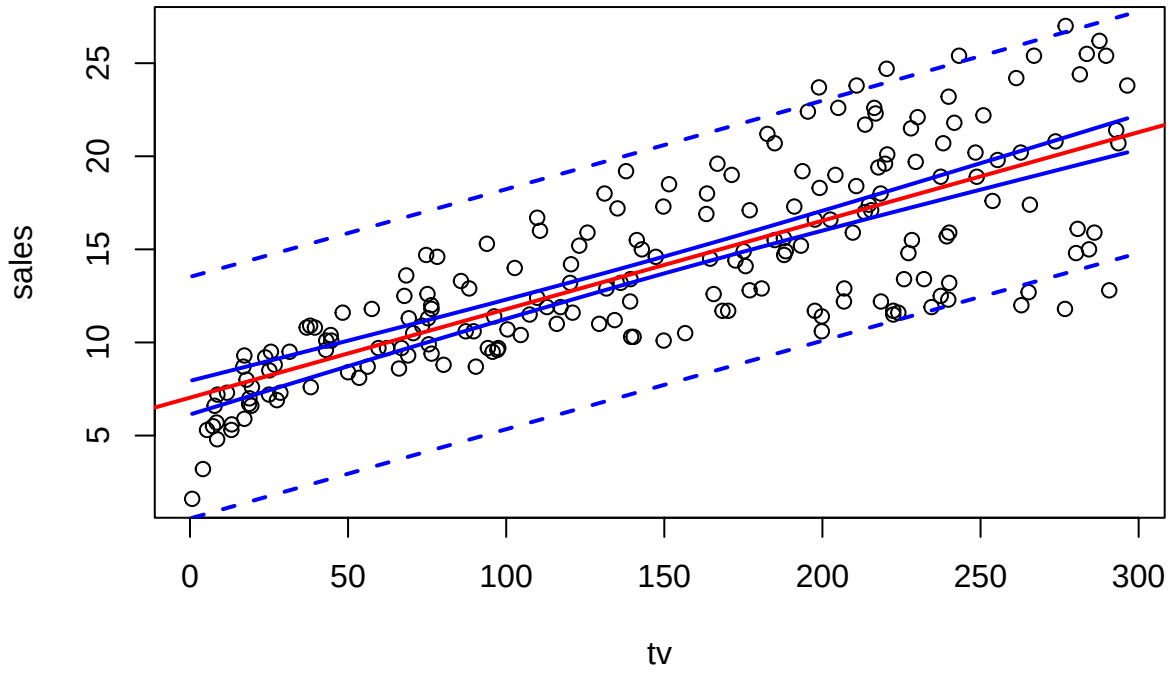
1.4 신뢰구간

앞선 분석결과를 통해 90% 신뢰구간을 구한다.

```
confint(lm.fit, level=0.90)
```

```
##                    5 %      95 %
## (Intercept) 6.27596881 7.7892183
## tv          0.04309018 0.0519831
```

```
#help(predict.lm)
#predict.lm(lm.fit,interval="confidence")
#predict.lm(lm.fit,interval="prediction")
attach(adv)
c.pred=predict(lm.fit, level = 0.95, interval="confidence")
p.pred=predict(lm.fit, level = 0.95, interval="prediction")
plot(tv,sales)
abline(lm.fit,col="red", lwd = 2)
o = order(tv, decreasing = F)
lines(tv[o],p.pred[,2][o], lty = "dashed", col="blue", type = "l", lwd = 2)
lines(tv[o],p.pred[,3][o], lty = "dashed", col="blue", lwd = 2)
lines(tv[o],c.pred[,2][o], col="blue", lwd = 2)
lines(tv[o],c.pred[,3][o], col="blue", lwd = 2)
```




```
detach(adv)
```

1.5 예측

독립변수가 새로운 수치일 경우에, sales의 수치를 예측한다.

```
predict(lm.fit, data.frame(tv=c(147)), level=0.90, interval="prediction")
```

```
##        fit      lwr      upr
## 1 14.02048 8.621824 19.41914
```

```r
predict(lm.fit, data.frame(tv=c(147)), level=0.90, interval="confidence")
```

```
##        fit      lwr      upr
## 1 14.02048 13.63969 14.40127
```

```r
predict(lm.fit, data.frame(tv=c(230.1, 44.5, 17.2)), level=0.95, interval="confidence")
```

```
##         fit       lwr       upr
## 1 17.970775 17.337774 18.603775
## 2  9.147974  8.439101  9.856848
## 3  7.850224  7.024932  8.675515
```

```r
predict(lm.fit, data.frame(tv=c(200, 50, 50)), level=0.90, interval="prediction")
```

```
##         fit       lwr      upr
## 1 16.539922 11.136133 21.94371
## 2  9.409426  3.993554 14.82530
## 3  9.409426  3.993554 14.82530
```

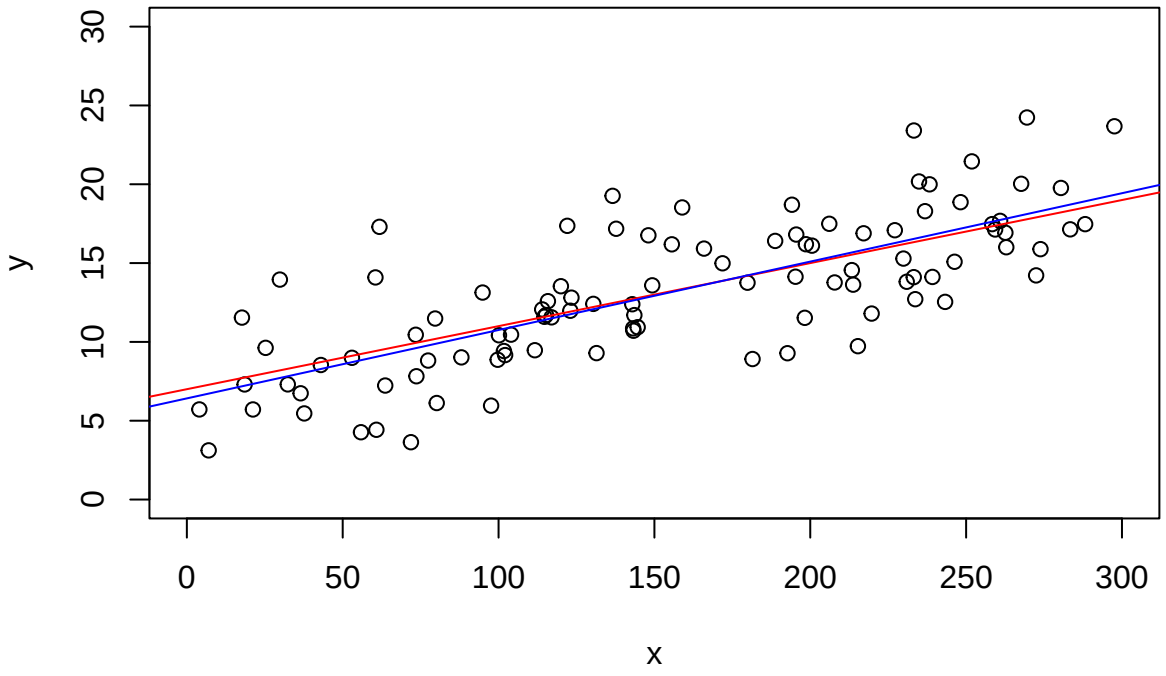
```r
predict(lm.fit, data.frame(tv=c(230.1, 44.5, 17.2)), level=0.95, interval="none")
```

```
##         1         2         3
## 17.970775  9.147974  7.850224
```

1.6 추정치의 변동

아래는 빨간색 회귀직선을 따르는 데이터를 생성한 후, 이 데이터에 회귀직선을 적합해서 그린 그림이다. 여러번 반복해보자.

```
x = runif(100)*300
y = 7.0 + 0.04*x + rnorm(100, 0, 3.259)

sim.fit = lm(y~x)

plot(x, y, xlim=c(0,300), ylim=c(0,30))
abline(c(7.0, 0.04), col="red")
abline(sim.fit,col="blue")
```

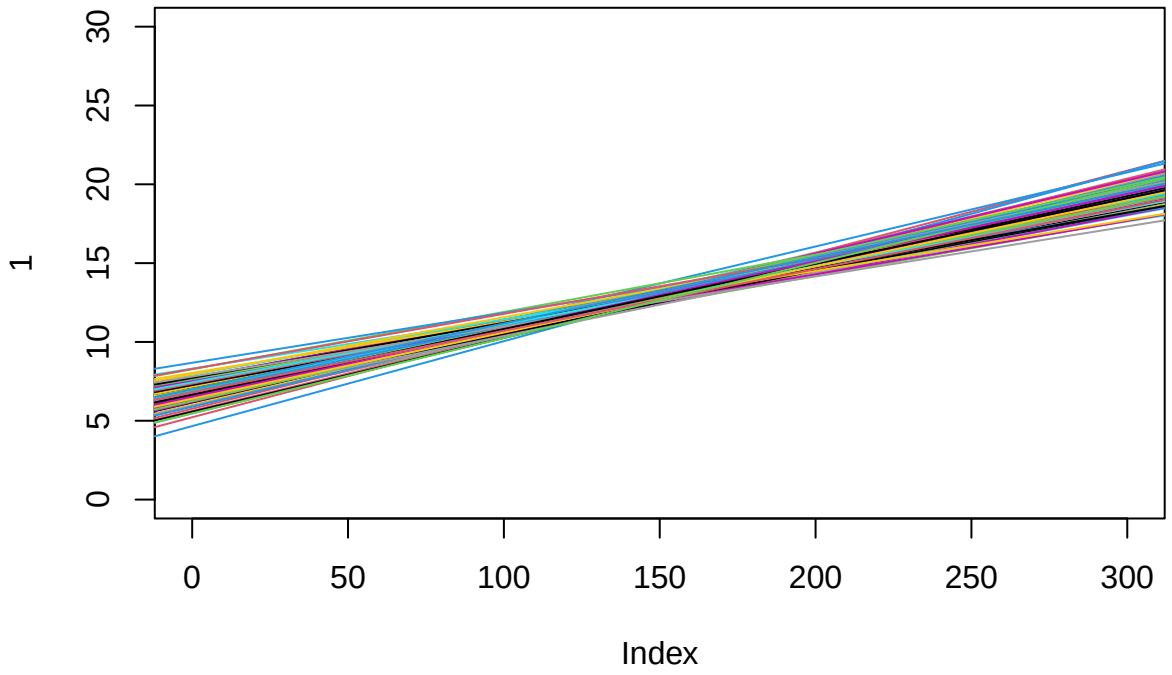



```
coef(sim.fit)
```

```
## (Intercept)           x
##  6.41557837  0.04339309
```

아래는 위의 작업을 100번 반복했다.

```
plot(1, xlim=c(0,300), ylim=c(0,30), type="n")
betas = numeric()
for(i in 1:100) {
  x = runif(100)*300
  y = 7.0 + 0.04*x + rnorm(100, 0, 3.259)
  sim.fit = lm(y~x)
  abline(sim.fit,col=i)
  betas = rbind(betas, coef(sim.fit))
}
```



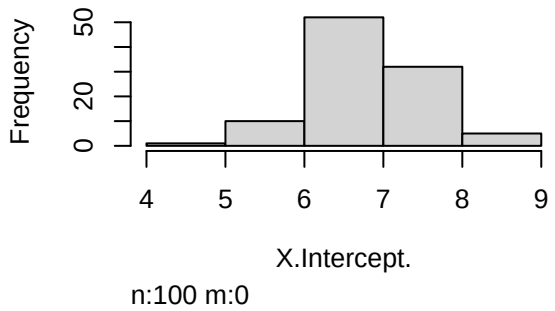

```
betas = data.frame(betas)
summary(betas)
```

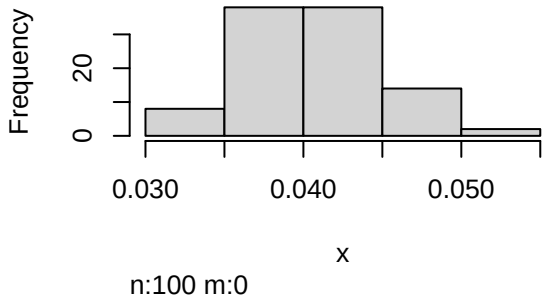
```
##   X.Intercept.          x
##  Min.   :4.661   Min.   :0.03114
##  1st Qu.:6.472   1st Qu.:0.03756
##  Median :6.876   Median :0.04044
##  Mean   :6.853   Mean   :0.04064
##  3rd Qu.:7.195   3rd Qu.:0.04285
##  Max.   :8.679   Max.   :0.05388
```

```
sapply(betas, sd)
```

```
## X.Intercept.            x
##  0.683286735  0.004240705
```

```
hist(betas)
```

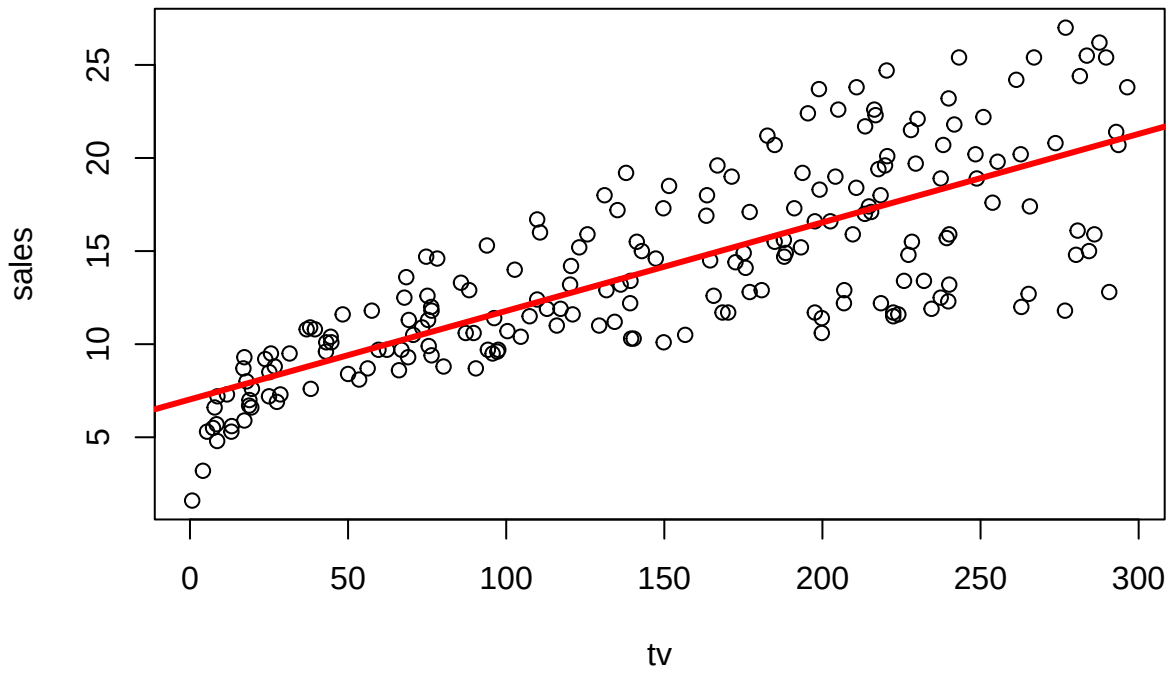




1.7 회귀직선 그리기

앞서 구한 회귀직선을 그린다.

```
attach(adv)
plot(tv, sales);abline(lm.fit, lwd=3, col="red")
```




```
detach(adv)
```

1.8 단순선형회귀 실습:

반응변수 sales, 설명변수 radio로 하는 단순선형회귀를 적합

회귀선을 추정하고, 추정치의 해석

자료의 산점도, fitted line, 그리고 prediction interval의 그림을 그린다.

1.9 다중회귀모형의 적합

반응변수를 sales, 독립변수를 tv, radio, newspaper로 하는 다중회귀분석을 진행한다.

```
lm.fit = lm(sales ~ tv + radio + newspaper, data=adv)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = sales ~ tv + radio + newspaper, data = adv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.938889   0.311908   9.422   <2e-16 ***
## tv           0.045765   0.001395  32.809   <2e-16 ***
## radio        0.188530   0.008611  21.893   <2e-16 ***
## newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

회귀계수의 두 가지 해석

2. 다중회귀모형의 해석:Orthogonalization

보스톤 집값자료

```
data(Boston)
boston = Boston %>%
         select(crim, chas, rm, age, tax, black, lstat, medv) %>%
         mutate(chas = as.factor(chas))
head(boston)
```

```
##      crim chas    rm  age tax  black lstat medv
## 1 0.00632    0 6.575 65.2 296 396.90  4.98 24.0
## 2 0.02731    0 6.421 78.9 242 396.90  9.14 21.6
## 3 0.02729    0 7.185 61.1 242 392.83  4.03 34.7
## 4 0.03237    0 6.998 45.8 222 394.63  2.94 33.4
## 5 0.06905    0 7.147 54.2 222 396.90  5.33 36.2
## 6 0.02985    0 6.430 58.7 222 394.12  5.21 28.7
```

CRIM : 타운별 1인당 범죄율 CHAS : 찰스강에 대한 더미변수(강의 경계 : 1, 아니면 0) RM : 주택 1가구당 평균 방의 개수 AGE : 1940년에 이전에 건축된 소유주택의 비율 TAX : 10,000 달러 당 재산세율 B : 자치시 별 흑인 비율 LSTAT : 하위 계층 비율 MEDV: 집값 -> 반응변수

Estimation of regression coefficient β_3 based on full model

```
lm.fit = lm(medv ~ rm + tax + lstat, data = boston)
```

Estimation of regreesion coefficient β_3 based on orthogonalization

```
lm.fit_1 = lm(medv ~ rm + tax, data = boston)
resid_1 = lm.fit_1$residuals
```

```
lm.fit_2 = lm(lstat ~ rm + tax, data = boston)
resid_2 = lm.fit_2$residuals
```

```
data_resid = data.frame(res1 = resid_1, res2 = resid_2)
lm.fit_3 = lm(resid_1 ~ 0 + resid_2, data = data_resid)
```

Compare results

```
coeff_lstat_lm.fit = as.vector(lm.fit$coefficients[4])
coeff_lstat_lm.fit_3 = as.vector(lm.fit_3$coefficients)
all.equal(coeff_lstat_lm.fit, coeff_lstat_lm.fit_3)
```

```
## [1] TRUE
```

Causation과 Association 차이를 소개한다.

3. 신용카드자료

이번엔 신용카드자료를 사용하여 다중회귀 분석을 진행한다.

3.1 자료 읽기

```
credit = read.csv("Credit.csv", header=T, sep=",")
credit = credit[,-1]
names(credit) = tolower(names(credit))
head(credit)
```

```
##    income limit rating cards age education gender student married ethnicity
## 1  14.891  3606    283     2  34        11   Male      No     Yes Caucasian
## 2 106.025  6645    483     3  82        15 Female     Yes     Yes     Asian
## 3 104.593  7075    514     4  71        11   Male      No      No     Asian
## 4 148.924  9504    681     3  36        11 Female      No      No     Asian
## 5  55.882  4897    357     2  68        16   Male      No     Yes Caucasian
## 6  80.180  8047    569     4  77        10   Male      No      No Caucasian
##   balance
## 1     333
## 2     903
## 3     580
## 4     964
## 5     331
## 6    1151
```

3.2 자료 탐색

다음의 코드들을 통해 탐색적 자료분석을 진행한다.

```
attach(credit)
library(Hmisc)
summary(credit)
```

```
##      income           limit           rating          cards      
##  Min.   : 10.35   Min.   :  855   Min.   : 93.0   Min.   :1.000  
##  1st Qu.: 21.01   1st Qu.: 3088   1st Qu.:247.2   1st Qu.:2.000  
##  Median : 33.12   Median : 4622   Median :344.0   Median :3.000  
##  Mean   : 45.22   Mean   : 4736   Mean   :354.9   Mean   :2.958  
##  3rd Qu.: 57.47   3rd Qu.: 5873   3rd Qu.:437.2   3rd Qu.:4.000  
##  Max.   :186.63   Max.   :13913   Max.   :982.0   Max.   :9.000  
##       age          education        gender            student     
##  Min.   :23.00   Min.   : 5.00   Length:400         Length:400        
##  1st Qu.:41.75   1st Qu.:11.00   Class :character   Class :character  
##  Median :56.00   Median :14.00   Mode  :character   Mode  :character  
##  Mean   :55.67   Mean   :13.45                                        
##  3rd Qu.:70.00   3rd Qu.:16.00                                        
##  Max.   :98.00   Max.   :20.00                                        
##    married           ethnicity           balance       
##  Length:400         Length:400         Min.   :   0.00  
##  Class :character   Class :character   1st Qu.:  68.75  
##  Mode  :character   Mode  :character   Median : 459.50  
##                                        Mean   : 520.01  
##                                        3rd Qu.: 863.00
```

```
##                                   Max.   :1999.00
```

```
Hmisc::describe(credit)
```

```
## credit
##
##  11  Variables      400  Observations
## ---------------------------------------------------------------------------
## income
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      400        0      399        1    45.22    35.34    12.07    14.58
##      .25      .50      .75      .90      .95
##    21.01    33.12    57.47    92.45   124.35
##
## lowest :  10.354  10.363  10.403  10.503  10.588
## highest: 163.329 180.379 180.682 182.728 186.634
## ---------------------------------------------------------------------------
## limit
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      400        0      387        1     4736     2546     1483     1919
##      .25      .50      .75      .90      .95
##     3088     4622     5873     7660     9162
##
## lowest :   855   886   905   906  1134, highest: 11589 11966 12066 13414 13913
## ---------------------------------------------------------------------------
## rating
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      400        0      283        1    354.9    170.5    138.0    167.0
##      .25      .50      .75      .90      .95
##    247.2    344.0    437.2    549.5    642.7
##
## lowest :  93 103 112 115 117, highest: 817 828 832 949 982
## ---------------------------------------------------------------------------
## cards
##        n  missing distinct     Info     Mean      Gmd
##      400        0        9    0.946    2.958    1.486
##
## lowest : 1 2 3 4 5, highest: 5 6 7 8 9
##
## Value          1     2     3     4     5     6     7     8     9
## Frequency     51   115   111    72    34    11     4     1     1
## Proportion 0.128 0.288 0.278 0.180 0.085 0.028 0.010 0.002 0.002
## ---------------------------------------------------------------------------
## age
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      400        0       68        1    55.67     19.9    29.00    32.00
##      .25      .50      .75      .90      .95
##    41.75    56.00    70.00    79.10    82.00
##
## lowest : 23 24 25 26 27, highest: 86 87 89 91 98
## ---------------------------------------------------------------------------
## education
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      400        0       16    0.991    13.45    3.551        8        9
##      .25      .50      .75      .90      .95
```

```
##       11       14       16       17       18
##
## lowest :  5  6  7  8  9, highest: 16 17 18 19 20
##
## Value          5     6     7     8     9    10    11    12    13    14    15
## Frequency      1     5     8    14    25    24    33    37    38    48    49
## Proportion 0.002 0.013 0.020 0.035 0.062 0.060 0.082 0.092 0.095 0.120 0.122
##
## Value         16    17    18    19    20
## Frequency     50    34    22    10     2
## Proportion 0.125 0.085 0.055 0.025 0.005
## ---------------------------------------------------------------------------
## gender
##        n  missing distinct
##      400        0        2
##
## Value        Male Female
## Frequency     193    207
## Proportion  0.482  0.517
## ---------------------------------------------------------------------------
## student
##        n  missing distinct
##      400        0        2
##
## Value       No Yes
## Frequency  360  40
## Proportion 0.9 0.1
## ---------------------------------------------------------------------------
## married
##        n  missing distinct
##      400        0        2
##
## Value         No   Yes
## Frequency    155   245
## Proportion 0.388 0.613
## ---------------------------------------------------------------------------
## ethnicity
##        n  missing distinct
##      400        0        3
##
## Value      African American            Asian        Caucasian
## Frequency                99              102              199
## Proportion            0.248            0.255            0.498
## ---------------------------------------------------------------------------
## balance
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      400        0      284    0.989      520      517     0.00     0.00
##      .25      .50      .75      .90      .95
##    68.75   459.50   863.00  1151.40  1355.30
##
## lowest :    0    5    8   15   16, highest: 1677 1687 1779 1809 1999
## ---------------------------------------------------------------------------
```

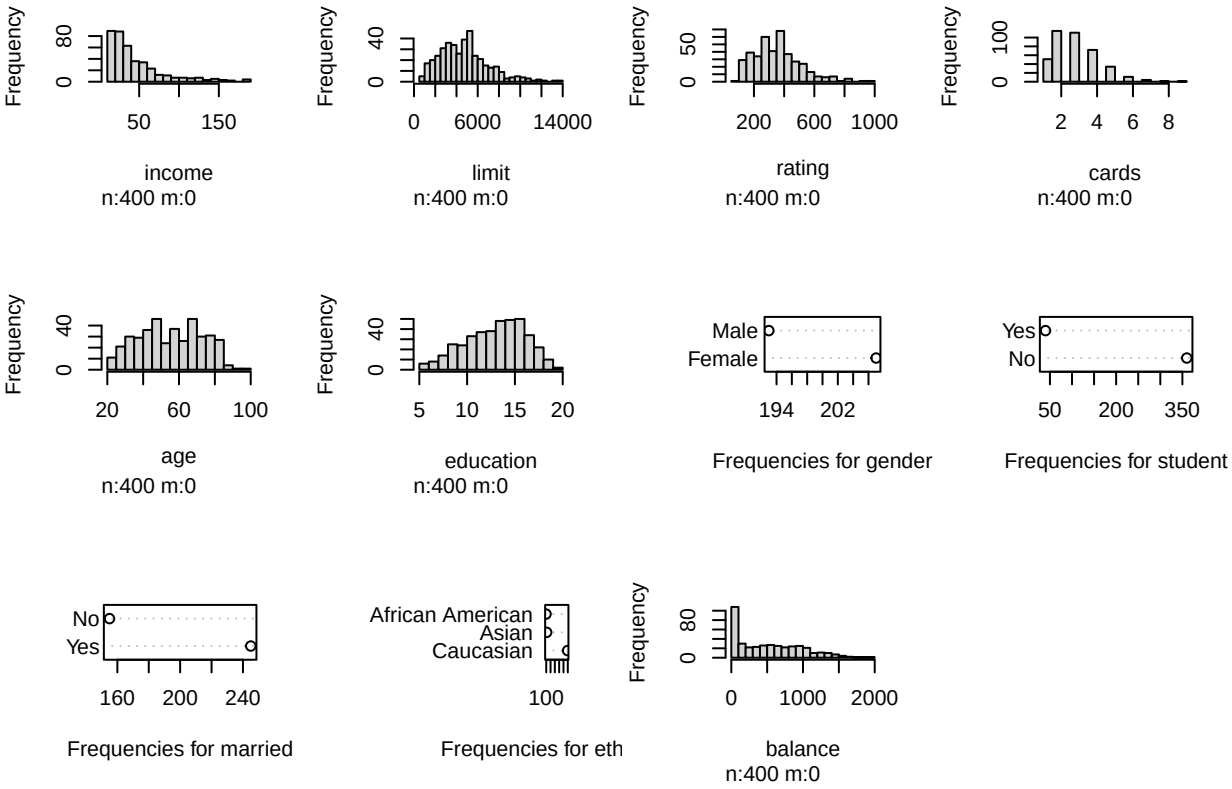
```
psych::describe(credit)
```

```
##            vars   n    mean      sd  median trimmed     mad    min      max
## income        1 400   45.22   35.24   33.12   38.67   23.12  10.35   186.63
## limit         2 400 4735.60 2308.20 4622.50 4557.31 2130.50 855.00 13913.00
## rating        3 400  354.94  154.72  344.00  342.14  140.85  93.00   982.00
## cards         4 400    2.96    1.37    3.00    2.87    1.48   1.00     9.00
## age           5 400   55.67   17.25   56.00   55.68   20.76  23.00    98.00
## education     6 400   13.45    3.13   14.00   13.57    2.97   5.00    20.00
## gender*       7 400    1.52    0.50    2.00    1.52    0.00   1.00     2.00
## student*      8 400    1.10    0.30    1.00    1.00    0.00   1.00     2.00
## married*      9 400    1.61    0.49    2.00    1.64    0.00   1.00     2.00
## ethnicity*   10 400    2.25    0.83    2.00    2.31    1.48   1.00     3.00
## balance      11 400  520.02  459.76  459.50  475.13  593.04   0.00  1999.00
##               range  skew kurtosis     se
## income       176.28  1.73     2.87   1.76
## limit      13058.00  0.83     0.96 115.41
## rating       889.00  0.86     1.01   7.74
## cards          8.00  0.79     0.90   0.07
## age           75.00  0.01    -1.08   0.86
## education     15.00 -0.33    -0.60   0.16
## gender*        1.00 -0.07    -2.00   0.03
## student*       1.00  2.66     5.07   0.02
## married*       1.00 -0.46    -1.79   0.02
## ethnicity*     2.00 -0.49    -1.37   0.04
## balance     1999.00  0.58    -0.55  22.99
```

```
sapply(credit[,-(7:10)], sd)
```

```
##      income       limit      rating       cards         age   education
##   35.244273 2308.198848  154.724143    1.371275   17.249807    3.125207
##     balance
##  459.758877
```

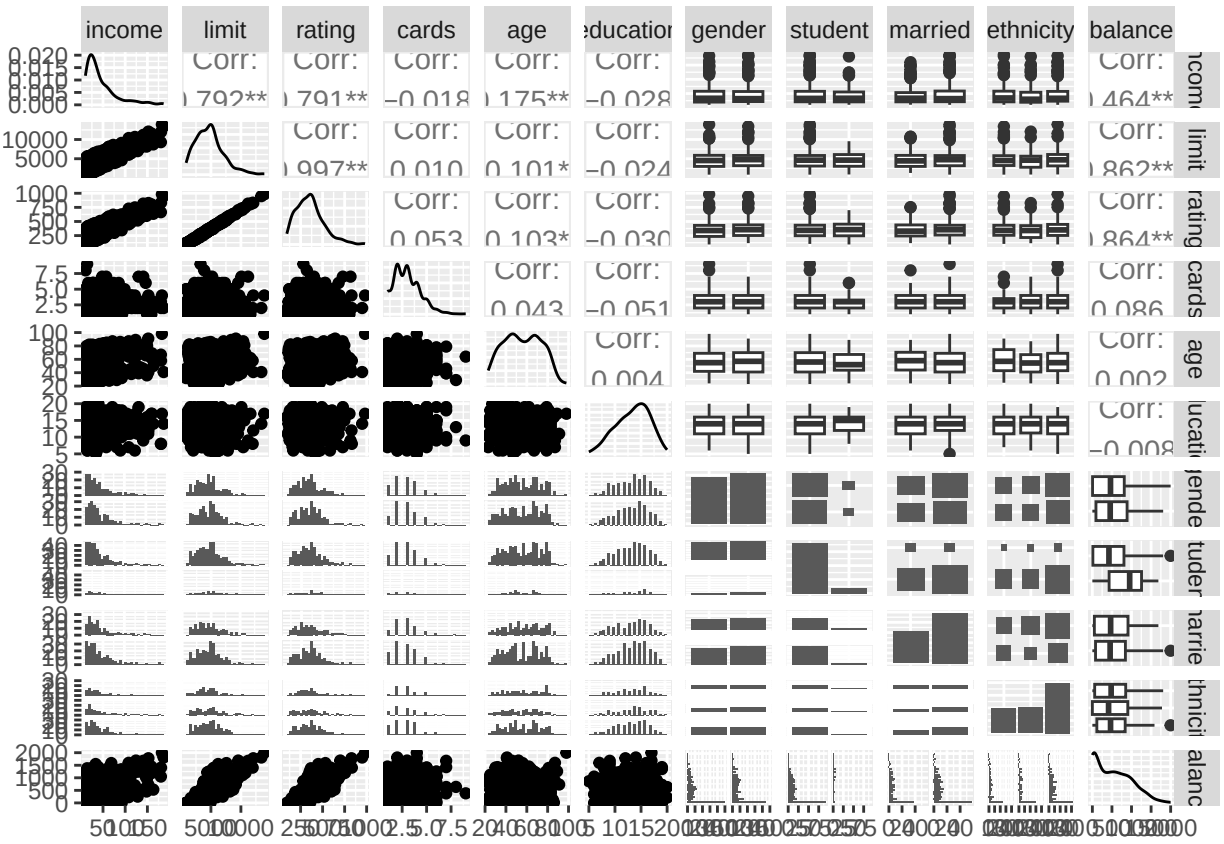
```
hist(credit)
```



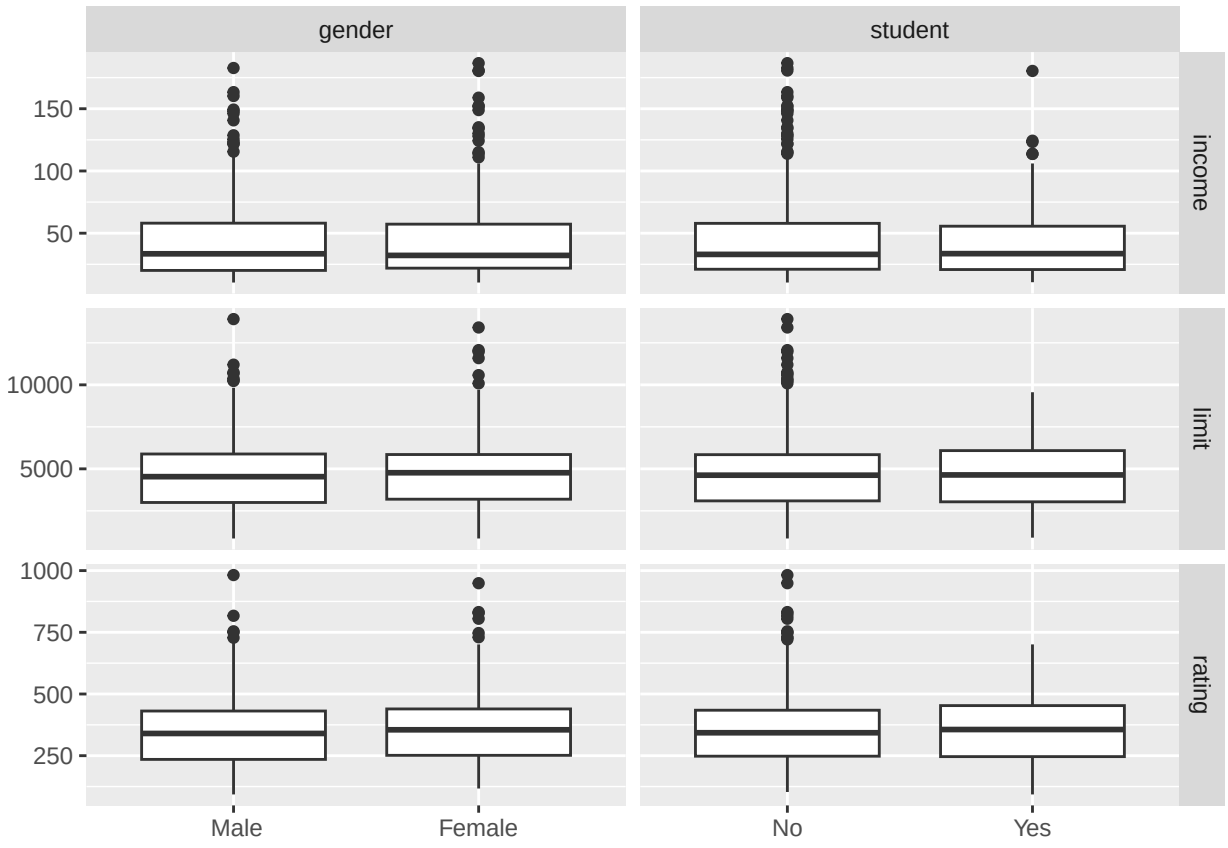

```
#pairs(credit)  #
cor(credit[,-(7:10)])
```

```
##                income       limit      rating       cards         age
## income     1.00000000  0.79208834  0.79137763 -0.01827261 0.175338403
## limit      0.79208834  1.00000000  0.99687974  0.01023133 0.100887922
## rating     0.79137763  0.99687974  1.00000000  0.05323903 0.103164996
## cards     -0.01827261  0.01023133  0.05323903  1.00000000 0.042948288
## age        0.17533840  0.10088792  0.10316500  0.04294829 1.000000000
## education -0.02769198 -0.02354853 -0.03013563 -0.05108422 0.003619285
## balance    0.46365646  0.86169727  0.86362516  0.08645635 0.001835119
##               education      balance
## income    -0.027691982  0.463656457
## limit     -0.023548534  0.861697267
## rating    -0.030135627  0.863625161
## cards     -0.051084217  0.086456347
## age        0.003619285  0.001835119
## education  1.000000000 -0.008061576
## balance   -0.008061576  1.000000000
```

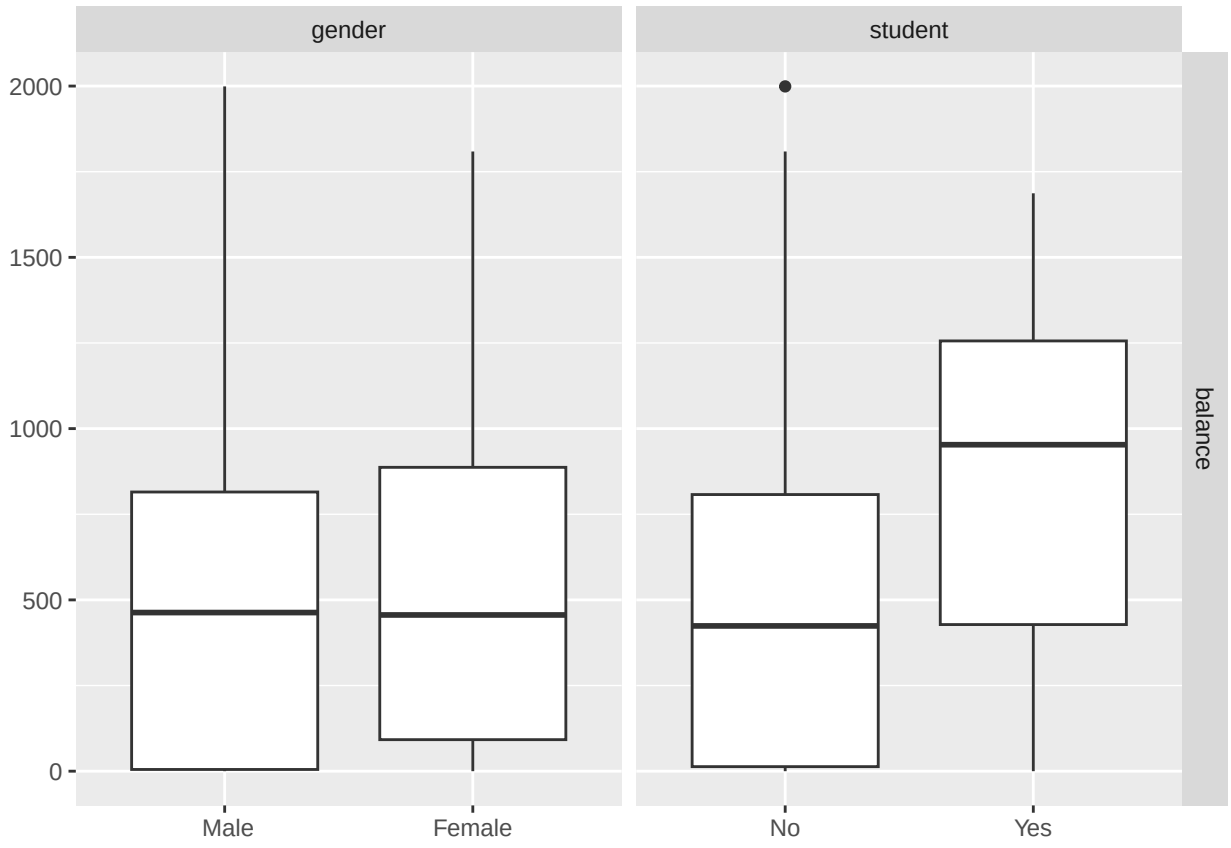
```
ggpairs(credit)
```




```
#
ggduo(credit, columnsX = 7:8, columnsY = 1:3)
```




```
ggduo(credit, columnsX = 7:8, columnsY = 11:11)
```




```
#
detach(credit)
```

3.3 가변수

독립변수가 범주형 변수일 경우에는 가변수를 사용하여 분석한다.

신용카드자료에서 gender와 ethnicity는 범주형 변수이며, 이를 사용하여 단순선형회귀분석을 진행한다.

```
attach(credit)

lm.fit1 = lm(balance ~ gender, data = credit)
summary(lm.fit1)
```

```
##
## Call:
## lm(formula = balance ~ gender, data = credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -529.54 -455.35  -60.17  334.71 1489.20
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    509.80      33.13  15.389   <2e-16 ***
## genderFemale    19.73      46.05   0.429    0.669
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 460.2 on 398 degrees of freedom
## Multiple R-squared:  0.0004611,  Adjusted R-squared:  -0.00205
## F-statistic: 0.1836 on 1 and 398 DF,  p-value: 0.6685
```

```
lm.fit2 = lm(balance ~ ethnicity, data = credit)
summary(lm.fit2)
```

```
##
## Call:
## lm(formula = balance ~ ethnicity, data = credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -531.00 -457.08  -63.25  339.25 1480.50
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           531.00      46.32  11.464   <2e-16 ***
## ethnicityAsian        -18.69      65.02  -0.287    0.774
## ethnicityCaucasian    -12.50      56.68  -0.221    0.826
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 460.9 on 397 degrees of freedom
## Multiple R-squared:  0.0002188,  Adjusted R-squared:  -0.004818
## F-statistic: 0.04344 on 2 and 397 DF,  p-value: 0.9575
```

```
fethnicity=as.factor(ethnicity)
str(fethnicity)
```

```
##  Factor w/ 3 levels "African American",..: 3 2 2 2 3 3 1 2 3 1 ...
```

```
detach(credit)
```

genderMale의 효과를 0으로 ethnicityBlakc의 효과를 0으로 가정한다.

3.4 중회귀모형의 적합

```
lm.fit3= lm(balance ~., data = credit)
summary(lm.fit3)
```

```
##
## Call:
## lm(formula = balance ~ ., data = credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -161.64  -77.70  -13.49   53.98  318.20
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -479.20787   35.77394 -13.395  < 2e-16 ***
## income               -7.80310    0.23423 -33.314  < 2e-16 ***
## limit                 0.19091    0.03278   5.824 1.21e-08 ***
## rating                1.13653    0.49089   2.315   0.0211 *
## cards                17.72448    4.34103   4.083 5.40e-05 ***
## age                  -0.61391    0.29399  -2.088   0.0374 *
## education            -1.09886    1.59795  -0.688   0.4921
## genderFemale        -10.65325    9.91400  -1.075   0.2832
## studentYes          425.74736   16.72258  25.459  < 2e-16 ***
## marriedYes           -8.53390   10.36287  -0.824   0.4107
## ethnicityAsian       16.80418   14.11906   1.190   0.2347
## ethnicityCaucasian   10.10703   12.20992   0.828   0.4083
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 98.79 on 388 degrees of freedom
## Multiple R-squared:  0.9551, Adjusted R-squared:  0.9538
## F-statistic: 750.3 on 11 and 388 DF,  p-value: < 2.2e-16
```

추정된 회귀식 (유의미한 변수들과 ethnicity 변수 포함 모형) 적어보자!!

4. 교호작용

4.1 광고 자료 (연속형*연속형)

```
# lm.fit = lm(sales ~ tv*radio, data=adv)
lm.fit = lm(sales ~ tv+radio+tv:radio, data=adv)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = sales ~ tv + radio + tv:radio, data = adv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3366 -0.4028  0.1831  0.5948  1.5246
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
## tv          1.910e-02  1.504e-03  12.699   <2e-16 ***
## radio       2.886e-02  8.905e-03   3.241   0.0014 **
## tv:radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9435 on 196 degrees of freedom
## Multiple R-squared:  0.9678, Adjusted R-squared:  0.9673
## F-statistic:  1963 on 3 and 196 DF,  p-value: < 2.2e-16
```

4.2 카시트 자료

```
data(Carseats)
head(Carseats)
```

```
##   Sales CompPrice Income Advertising Population Price ShelveLoc Age Education
## 1  9.50       138     73          11        276   120       Bad  42        17
## 2 11.22       111     48          16        260    83      Good  65        10
## 3 10.06       113     35          10        269    80    Medium  59        12
## 4  7.40       117    100           4        466    97    Medium  55        14
## 5  4.15       141     64           3        340   128       Bad  38        13
## 6 10.81       124    113          13        501    72       Bad  78        16
##   Urban  US
## 1   Yes Yes
## 2   Yes Yes
## 3   Yes Yes
## 4   Yes Yes
## 5   Yes  No
## 6    No Yes
```

“dplyr” 패키지:

filter() 지정한 조건식에 맞는 데이터 추출 subset()

select() 열의 추출 data[, c(“Year”, “Month”)]

mutate() 열 추가 transform()

arrange() 정렬 order(), sort()

summarise() 집계

4.3 범주형 vs 범주형

$$Sales = \beta_0 + \beta_1 USyes + \beta_2 Urbanyes + \beta_3 USyes * Urbanyes + \epsilon$$

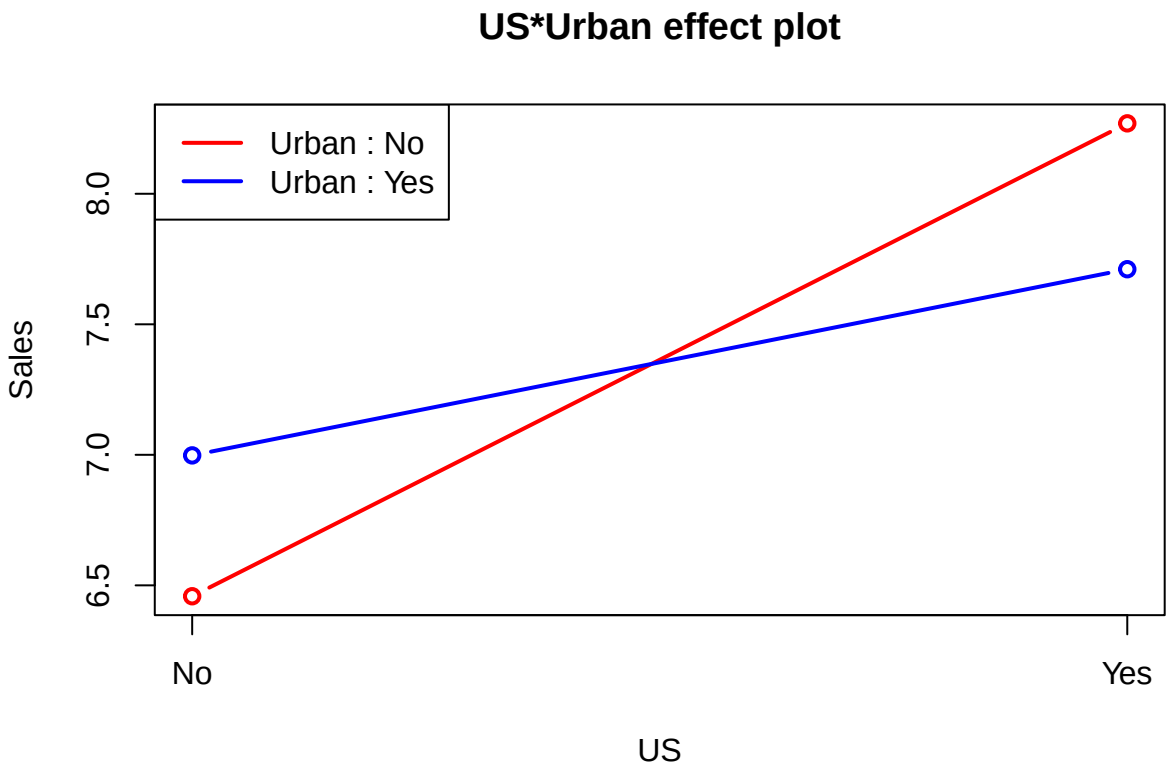
```
Carseats1= Carseats %>% select(Sales, Urban, US)
head(Carseats1)
```

```
##   Sales Urban  US
## 1  9.50   Yes Yes
## 2 11.22   Yes Yes
## 3 10.06   Yes Yes
## 4  7.40   Yes Yes
## 5  4.15   Yes  No
## 6 10.81    No Yes
```

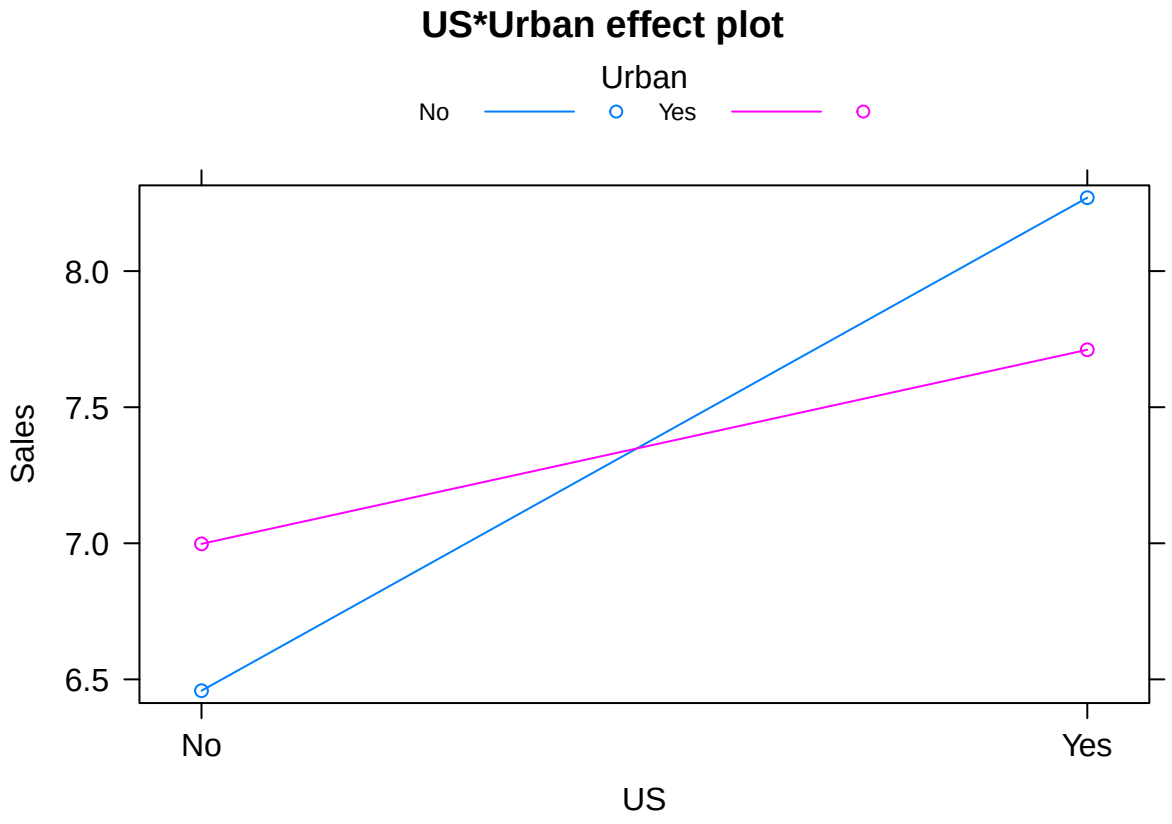
```
fit1 = lm(Sales ~ US * Urban, data = Carseats1)
summary(fit1)
```

```
##
## Call:
## lm(formula = Sales ~ US * Urban, data = Carseats1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.341 -1.961 -0.016  1.812  8.559
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.4583     0.4097  15.764  < 2e-16 ***
## USYes            1.8115     0.5245   3.454 0.000612 ***
## UrbanYes         0.5396     0.4982   1.083 0.279512
## USYes:UrbanYes  -1.0983     0.6301  -1.743 0.082081 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.779 on 396 degrees of freedom
## Multiple R-squared:  0.0393, Adjusted R-squared:  0.03202
## F-statistic:   5.4 on 3 and 396 DF,  p-value: 0.001191
```

```
coeff <- fit1$coefficients
x = c(0,1)

Urban_No = matrix(0,2,1)
Urban_No[1,1] <- t(coeff)%*%c(1,0,0,0) # beta_0
Urban_No[2,1] <- t(coeff)%*%c(1,1,0,0) # beta_0 + beta_1

Urban_Yes <- matrix(0,2,1)
Urban_Yes[1,1] <- t(coeff)%*%c(1,0,1,0) # beta_0 + beta_2
Urban_Yes[2,1] <- t(coeff)%*%c(1,1,1,1) # beta_0 + beta_1 + beta_2 + beta_3

plot(x, Urban_No, type = "b", col = "red", lwd = 2,
     xlab = "US", ylab = "Sales", main = "US*Urban effect plot", xaxt = "n")
axis(side = 1, at = c(0,1), labels = c("No", "Yes"))
lines(x, Urban_Yes, type = "b", col = "blue", lwd = 2)
legend("topleft", legend = c("Urban : No", "Urban : Yes"), col = c("red", "blue"), lwd = 2)
```




```
a = effect(term = "US*Urban", mod  = fit1)
plot(a, multiline =  TRUE)
```



4.4 범주형 vs 연속형

$$Sales = \beta_0 + \beta_1 Price + \beta_2 Urban + \beta_3 Urban * Price + \epsilon$$

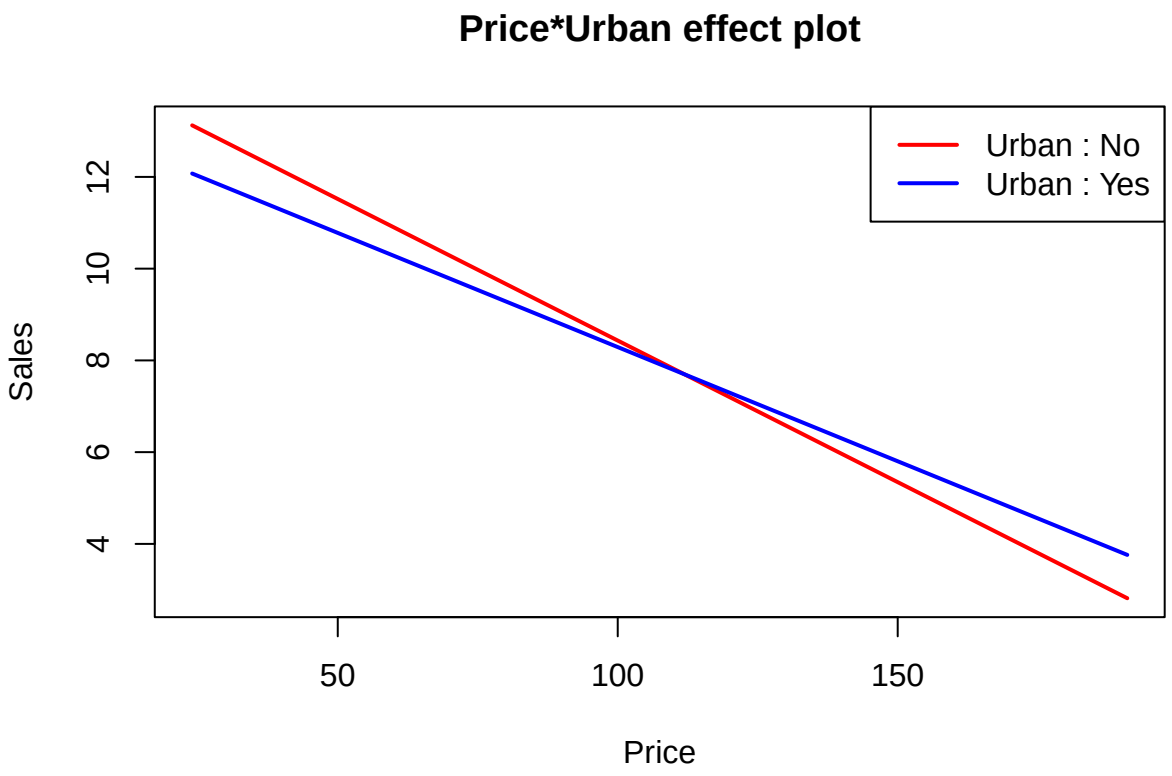
```
Carseats2 = Carseats %>% select(Sales, Price, Urban)
head(Carseats2)
```

```
##   Sales Price Urban
## 1  9.50   120   Yes
## 2 11.22    83   Yes
## 3 10.06    80   Yes
## 4  7.40    97   Yes
## 5  4.15   128   Yes
## 6 10.81    72    No
```

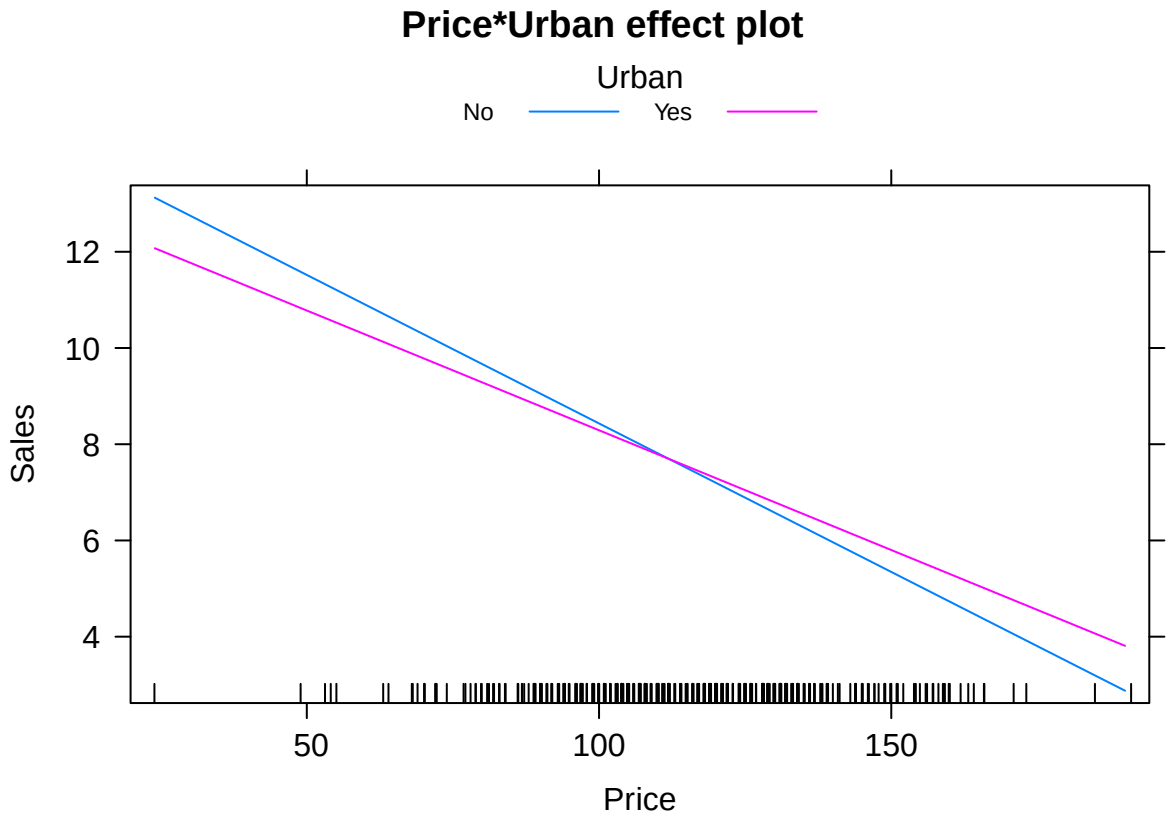
```
fit1 = lm(Sales ~ Price * Urban, data = Carseats2)
summary(fit1)
```

```
##
## Call:
## lm(formula = Sales ~ Price * Urban, data = Carseats2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5340 -1.8539 -0.0799  1.6758  7.5815
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    14.60526    1.18398  12.336  < 2e-16 ***
## Price          -0.06173    0.01018  -6.067 3.06e-09 ***
## UrbanYes       -1.33702    1.40228  -0.953    0.341
## Price:UrbanYes  0.01195    0.01198   0.998    0.319
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.535 on 396 degrees of freedom
## Multiple R-squared:    0.2,  Adjusted R-squared:  0.194
## F-statistic:    33 on 3 and 396 DF,  p-value: < 2.2e-16
```

```
coeff = fit1$coefficients
x = seq(min(Carseats2$Price), max(Carseats2$Price), length = 10) # grid points
# beta_0 + beta_1 * Price
Urban_No = Vectorize(function(x) t(coeff) %*% c(1, x, 0, 0))
# (beta_0 + beta_2) + (beta_1 + beta_3) * Price
Urban_Yes = Vectorize(function(x) t(coeff) %*% c(1, x, 1, x))

plot(x, Urban_No(x), type = "l", col = "red", lwd = 2,
     xlab = "Price", ylab = "Sales", main = "Price*Urban effect plot")
lines(x, Urban_Yes(x), type = "l", col = "blue", lwd = 2)
legend("topright", legend = c("Urban : No", "Urban : Yes"), col = c("red", "blue"), lwd = 2)
```




```
a = effect(term = "Price*Urban", mod = fit1)
plot(a, multiline = TRUE)
```



5. 실습: 보스톤 집값자료

```
data(Boston)
boston = Boston %>%
         select(crim, chas, rm, age, tax, black, lstat, medv) %>%
         mutate(chas = as.factor(chas))
head(Boston)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
##   medv
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

CRIM : 타운별 1인당 범죄율 CHAS : 찰스강에 대한 더미변수(강의 경계 : 1, 아니면 0) RM : 주택 1가구당 평균 방의 개수 AGE : 1940년에 이전에 건축된 소유주택의 비율 TAX : 10,000 달러 당 재산세율 B : 자치시 별 흑인 비율 LSTAT : 하위 계층 비율 MEDV: 집값 -> 반응변수

교호작용이 없는 중회귀모형을 적합하여라

교호작용이 있는 중회귀 모형을 적합하고 유의미한 교호작용을 찾아 해석하여라

6. 변수선택(optional)

변수선택의 절차는 다음과 같다. 전진선택법 .

1. 변수선텍의 기준을 정한다. adjusted R^2, Mallow's C_p, 여러 information criteria
2. 현재 모형에서 변수 하나를 추가하였을 때 adjuated R^2를 가장 높여주는 변수를 선택한다.
3. 선택된 변수를 추가하며 생기는 추가제곱합에 대한 F-검정을 실시한다.

3.1 추가제곱합에 대한 F-검정이 유의하면 추가를 유지하고 2.의 절차를 다시 진행한다.

3.2 추가제곱합에 대한 F-검정이 유의하지 않으면 선택된 변수를 모형에 포함시키지 않고 절차를 멈춘다.

```r
hitters.dat = read.csv("Hitters.csv") %>% na.omit() %>% select(AtBat:CWalks, PutOuts:Salary)
head(hitters.dat)
```

```
##   AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun CRuns CRBI CWalks
## 2   315   81     7   24  38    39    14   3449   835     69   321  414    375
## 3   479  130    18   66  72    76     3   1624   457     63   224  266    263
## 4   496  141    20   65  78    37    11   5628  1575    225   828  838    354
## 5   321   87    10   39  42    30     2    396   101     12    48   46     33
## 6   594  169     4   74  51    35    11   4408  1133     19   501  336    194
## 7   185   37     1   23   8    21     2    214    42      1    30    9     24
##   PutOuts Assists Errors Salary
## 2     632      43     10  475.0
## 3     880      82     14  480.0
## 4     200      11      3  500.0
## 5     805      40      4   91.5
## 6     282     421     25  750.0
## 7      76     127      7   70.0
```

```r
lm.fit = lm(Salary ~., data = hitters.dat)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = Salary ~ ., data = hitters.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -982.81 -187.84  -35.66  130.61 1947.43
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 126.10553   83.62448   1.508 0.132838
## AtBat        -2.20302    0.63605  -3.464 0.000629 ***
## Hits          7.82776    2.40198   3.259 0.001276 **
## HmRun         2.16355    6.23618   0.347 0.728937
## Runs         -2.09957    3.00849  -0.698 0.485911
## RBI          -0.02292    2.61033  -0.009 0.993003
## Walks         6.15106    1.84028   3.342 0.000960 ***
## Years        -2.59237   12.45401  -0.208 0.835280
## CAtBat       -0.17628    0.13667  -1.290 0.198325
## CHits         0.06976    0.67874   0.103 0.918221
## CHmRun       -0.23309    1.63561  -0.143 0.886795
## CRuns         1.61005    0.75162   2.142 0.033168 *
## CRBI          0.80143    0.70000   1.145 0.253367
## CWalks       -0.79394    0.33243  -2.388 0.017681 *
## PutOuts       0.29457    0.07830   3.762 0.000211 ***
## Assists       0.38400    0.22383   1.716 0.087499 .
## Errors       -2.87871    4.42077  -0.651 0.515539
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 319.9 on 246 degrees of freedom
## Multiple R-squared:  0.5279, Adjusted R-squared:  0.4972
## F-statistic: 17.19 on 16 and 246 DF,  p-value: < 2.2e-16
```

```r
# based on p-value
ols_step_forward_p(lm.fit, penter = 0.05, progress = T, details = F)
```

```
## Forward Selection Method
## ---------------------------
##
## Candidate Terms:
##
## 1. AtBat
## 2. Hits
## 3. HmRun
## 4. Runs
## 5. RBI
## 6. Walks
## 7. Years
## 8. CAtBat
## 9. CHits
## 10. CHmRun
## 11. CRuns
## 12. CRBI
## 13. CWalks
## 14. PutOuts
## 15. Assists
## 16. Errors
##
## We are selecting variables based on p value...
##
## Variables Entered:
##
## - CRBI
## - Hits
## - PutOuts
## - AtBat
## - Walks
##
## No more variables to be added.
##
## Final Model Output
## ------------------
##
##                            Model Summary
## ------------------------------------------------------------------
## R                       0.700       RMSE                  325.148
## R-Squared               0.490       Coef. Var              60.670
## Adj. R-Squared          0.481       MSE                105721.491
## Pred R-Squared          0.451       MAE                   218.843
## ------------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                  ANOVA
## --------------------------------------------------------------------------
##                     Sum of
```

```
##                    Squares         DF    Mean Square      F         Sig.
## ----------------------------------------------------------------------------
## Regression    26148689.566          5    5229737.913    49.467    0.0000
## Residual      27170423.223        257     105721.491
## Total         53319112.789        262
## ----------------------------------------------------------------------------
##
##                                  Parameter Estimates
## ----------------------------------------------------------------------------------------------
##       model      Beta    Std. Error    Std. Beta      t        Sig       lower      upper
## ----------------------------------------------------------------------------------------------
## (Intercept)    25.282        62.369                0.405    0.686    -97.538    148.102
##        CRBI     0.642         0.065        0.460    9.799    0.000      0.513      0.771
##        Hits     8.184         1.679        0.819    4.874    0.000      4.878     11.491
##     PutOuts     0.265         0.076        0.164    3.481    0.001      0.115      0.414
##       AtBat    -2.035         0.533       -0.665   -3.816    0.000     -3.085     -0.985
##       Walks     3.906         1.228        0.188    3.180    0.002      1.487      6.325
## ----------------------------------------------------------------------------------------------

##
##                              Selection Summary
## ------------------------------------------------------------------------------
##         Variable                    Adj.
## Step    Entered     R-Square    R-Square     C(p)          AIC          RMSE
## ------------------------------------------------------------------------------
##    1    CRBI          0.3215      0.3189    94.5992    3864.1393    372.3163
##    2    Hits          0.4252      0.4208    42.5217    3822.4873    343.3240
##    3    PutOuts       0.4514      0.4451    30.8657    3812.2144    336.0530
##    4    AtBat         0.4704      0.4622    22.9952    3804.9730    330.8397
##    5    Walks         0.4904      0.4805    14.5480    3796.8244    325.1484
## ------------------------------------------------------------------------------
```

```
ols_step_backward_p(lm.fit, prem = 0.05, progress = T, details = F)
```

```
## Backward Elimination Method
## ---------------------------
##
## Candidate Terms:
##
## 1 . AtBat
## 2 . Hits
## 3 . HmRun
## 4 . Runs
## 5 . RBI
## 6 . Walks
## 7 . Years
## 8 . CAtBat
## 9 . CHits
## 10 . CHmRun
## 11 . CRuns
## 12 . CRBI
## 13 . CWalks
## 14 . PutOuts
## 15 . Assists
## 16 . Errors
```

```
##
## We are eliminating variables based on p value...
##
## Variables Removed:
##
## - RBI
## - CHits
## - Years
## - CHmRun
## - HmRun
## - Errors
## - Runs
## - Assists
##
## No more variables satisfy the condition of p value = 0.05
##
##
## Final Model Output
## ------------------
##
##                            Model Summary
## ----------------------------------------------------------------------
## R                       0.720       RMSE                   317.822
## R-Squared               0.519       Coef. Var               59.303
## Adj. R-Squared          0.504       MSE                 101010.931
## Pred R-Squared          0.456       MAE                    215.946
## ----------------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                  ANOVA
## --------------------------------------------------------------------------
##                     Sum of
##                    Squares         DF    Mean Square      F         Sig.
## --------------------------------------------------------------------------
## Regression    27662336.204          8    3457792.026    34.232    0.0000
## Residual      25656776.584        254     101010.931
## Total         53319112.789        262
## --------------------------------------------------------------------------
##
##                                  Parameter Estimates
## ----------------------------------------------------------------------------------------------
##       model      Beta    Std. Error    Std. Beta      t        Sig       lower      upper
## ----------------------------------------------------------------------------------------------
## (Intercept)    88.369        64.530                 1.369    0.172    -38.712    215.450
##       AtBat    -2.064         0.529       -0.674    -3.903    0.000     -3.105     -1.022
##        Hits     7.157         1.663        0.716     4.304    0.000      3.882     10.431
##       Walks     5.478         1.610        0.264     3.402    0.001      2.307      8.650
##      CAtBat    -0.117         0.053       -0.595    -2.193    0.029     -0.223     -0.012
##       CRuns     1.370         0.386        1.006     3.553    0.000      0.610      2.129
##        CRBI     0.648         0.204        0.464     3.179    0.002      0.246      1.049
##      CWalks    -0.750         0.267       -0.439    -2.809    0.005     -1.275     -0.224
##     PutOuts     0.280         0.075        0.174     3.737    0.000      0.133      0.428
```

```
## ----------------------------------------------------------------------------------------
##
##
##                              Elimination Summary
## ------------------------------------------------------------------------
##         Variable                 Adj.
## Step    Removed     R-Square    R-Square     C(p)         AIC         RMSE
## ------------------------------------------------------------------------
##    1    RBI           0.5279      0.4993    15.0001    3796.7144    319.2242
##    2    CHits         0.5279      0.5013    13.0112    3794.7263    318.5872
##    3    Years         0.5278      0.5032    11.0643    3792.7831    317.9811
##    4    CHmRun        0.5276      0.5049     9.1650    3790.8907    317.4095
##    5    HmRun         0.5272      0.5064     7.3964    3789.1378    316.9254
##    6    Errors        0.5265      0.5078     5.7242    3787.4875    316.5063
##    7    Runs          0.5255      0.5086     4.2565    3786.0543    316.2207
##    8    Assists       0.5188      0.5037     5.7544    3787.7489    317.8222
## ------------------------------------------------------------------------
```

```r
# based on AIC
ols_step_forward_aic(lm.fit, progress = T, details = F)
```

```
## Forward Selection Method
## ------------------------
##
## Candidate Terms:
##
## 1 . AtBat
## 2 . Hits
## 3 . HmRun
## 4 . Runs
## 5 . RBI
## 6 . Walks
## 7 . Years
## 8 . CAtBat
## 9 . CHits
## 10 . CHmRun
## 11 . CRuns
## 12 . CRBI
## 13 . CWalks
## 14 . PutOuts
## 15 . Assists
## 16 . Errors
##
##
## Variables Entered:
##
## - CRBI
## - Hits
## - PutOuts
## - AtBat
## - Walks
##
## No more variables to be added.
##
## Final Model Output
## ------------------
##
##                            Model Summary
## ------------------------------------------------------------------
## R                       0.700       RMSE                  325.148
## R-Squared               0.490       Coef. Var              60.670
## Adj. R-Squared          0.481       MSE                105721.491
## Pred R-Squared          0.451       MAE                   218.843
## ------------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                 ANOVA
## --------------------------------------------------------------------------
##                     Sum of
##                    Squares         DF    Mean Square      F         Sig.
```

```
## ----------------------------------------------------------------------
## Regression    26148689.566          5    5229737.913    49.467    0.0000
## Residual      27170423.223        257     105721.491
## Total         53319112.789        262
## ----------------------------------------------------------------------
##
##                                 Parameter Estimates
## ------------------------------------------------------------------------------------------
##       model      Beta    Std. Error    Std. Beta      t        Sig       lower      upper
## ------------------------------------------------------------------------------------------
## (Intercept)    25.282        62.369                0.405    0.686    -97.538    148.102
##        CRBI     0.642         0.065        0.460    9.799    0.000      0.513      0.771
##        Hits     8.184         1.679        0.819    4.874    0.000      4.878     11.491
##     PutOuts     0.265         0.076        0.164    3.481    0.001      0.115      0.414
##       AtBat    -2.035         0.533       -0.665   -3.816    0.000     -3.085     -0.985
##       Walks     3.906         1.228        0.188    3.180    0.002      1.487      6.325
## ------------------------------------------------------------------------------------------

##
##                              Selection Summary
## ------------------------------------------------------------------------------
## Variable       AIC          Sum Sq           RSS          R-Sq      Adj. R-Sq
## ------------------------------------------------------------------------------
## CRBI         3864.139    17139433.534    36179679.255    0.32145      0.31885
## Hits         3822.487    22672552.898    30646559.890    0.42522      0.42080
## PutOuts      3812.214    24069815.933    29249296.856    0.45143      0.44508
## AtBat        3804.973    25079748.980    28239363.809    0.47037      0.46216
## Walks        3796.824    26148689.566    27170423.223    0.49042      0.48050
## ------------------------------------------------------------------------------
```

```
ols_step_backward_aic(lm.fit, progress = T, details = F)
```

```
## Backward Elimination Method
## ---------------------------
##
## Candidate Terms:
##
## 1 . AtBat
## 2 . Hits
## 3 . HmRun
## 4 . Runs
## 5 . RBI
## 6 . Walks
## 7 . Years
## 8 . CAtBat
## 9 . CHits
## 10 . CHmRun
## 11 . CRuns
## 12 . CRBI
## 13 . CWalks
## 14 . PutOuts
## 15 . Assists
## 16 . Errors
##
##
```

```
## Variables Removed:
##
## - RBI
## - CHits
## - Years
## - CHmRun
## - HmRun
## - Errors
## - Runs
##
## No more variables to be removed.
##
## Final Model Output
## ------------------
##
##                            Model Summary
## ------------------------------------------------------------------
## R                        0.725       RMSE                 316.221
## R-Squared                0.526       Coef. Var             59.005
## Adj. R-Squared           0.509       MSE                99995.553
## Pred R-Squared           0.457       MAE                  217.400
## ------------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                 ANOVA
## --------------------------------------------------------------------------
##                    Sum of
##                   Squares         DF    Mean Square      F         Sig.
## --------------------------------------------------------------------------
## Regression    28020237.892          9    3113359.766    31.135    0.0000
## Residual      25298874.897        253      99995.553
## Total         53319112.789        262
## --------------------------------------------------------------------------
##
##                                  Parameter Estimates
## ----------------------------------------------------------------------------------------------
##       model       Beta    Std. Error    Std. Beta      t        Sig       lower      upper
## ----------------------------------------------------------------------------------------------
## (Intercept)    107.341        64.983                 1.652    0.100    -20.635    235.317
##       AtBat     -2.308         0.542       -0.753    -4.260    0.000     -3.374     -1.241
##        Hits      7.422         1.660        0.742     4.470    0.000      4.152     10.692
##       Walks      5.716         1.607        0.275     3.557    0.000      2.551      8.881
##      CAtBat     -0.149         0.056       -0.757    -2.674    0.008     -0.259     -0.039
##       CRuns      1.535         0.393        1.127     3.901    0.000      0.760      2.309
##        CRBI      0.768         0.213        0.551     3.615    0.000      0.350      1.187
##      CWalks     -0.807         0.267       -0.472    -3.020    0.003     -1.333     -0.281
##     PutOuts      0.301         0.075        0.187     3.992    0.000      0.153      0.450
##     Assists      0.302         0.160        0.097     1.892    0.060     -0.012      0.617
## ----------------------------------------------------------------------------------------------
##
##
```

```
##                          Backward Elimination Summary
## ------------------------------------------------------------------------------
## Variable        AIC           RSS            Sum Sq        R-Sq      Adj. R-Sq
## ------------------------------------------------------------------------------
## Full Model    3798.714    25170309.440    28148803.349    0.52793      0.49723
## RBI           3796.714    25170317.325    28148795.463    0.52793      0.49926
## CHits         3794.726    25171450.927    28147661.862    0.52791      0.50126
## Years         3792.783    25176890.419    28142222.370    0.52781      0.50315
## CHmRun        3790.891    25187193.291    28131919.497    0.52761      0.50494
## HmRun         3789.138    25210866.819    28108245.969    0.52717      0.50645
## Errors        3787.487    25244408.197    28074704.591    0.52654      0.50775
## Runs          3786.054    25298874.897    28020237.892    0.52552      0.50864
## ------------------------------------------------------------------------------
```