`4.` **R 실습**. 보스턴 집값 데이터 (데이터 출처 : MASS 패키지). 이 데이터는 Boston 근처 지역의 지역적 특징과 주택 가격의 중앙값 등을 포함하고 있다. 데이터는 MASS 패키지 설치를 통해 Boston 데이터를 사용할 수 있다. 아래와 같이 사용가능하며 자세한 내용을 살펴볼 수 있다.

```
library(MASS)
head(Boston)
?Boston # Boston 데이터의 자세한 설명을 볼 수 있음
```

1인당 범죄율 `crim` 을 설명변수 x로 하고, 주택가격의 중앙값 `medv` 을 반응변수 y 로 할 때 다음에 대하여 답하시오.

In [3]:
```
### 라이브러리 불러오기
##library(MASS)
##library(lmtest)
```

(1) 선형 회귀모형 $y = \beta_0 + \beta_1 x + \epsilon$을 적합시켜라.

In [4]:
```
x <- Boston$crim
y <- Boston$medv

model <- lm(y~x)

summary(model)
```

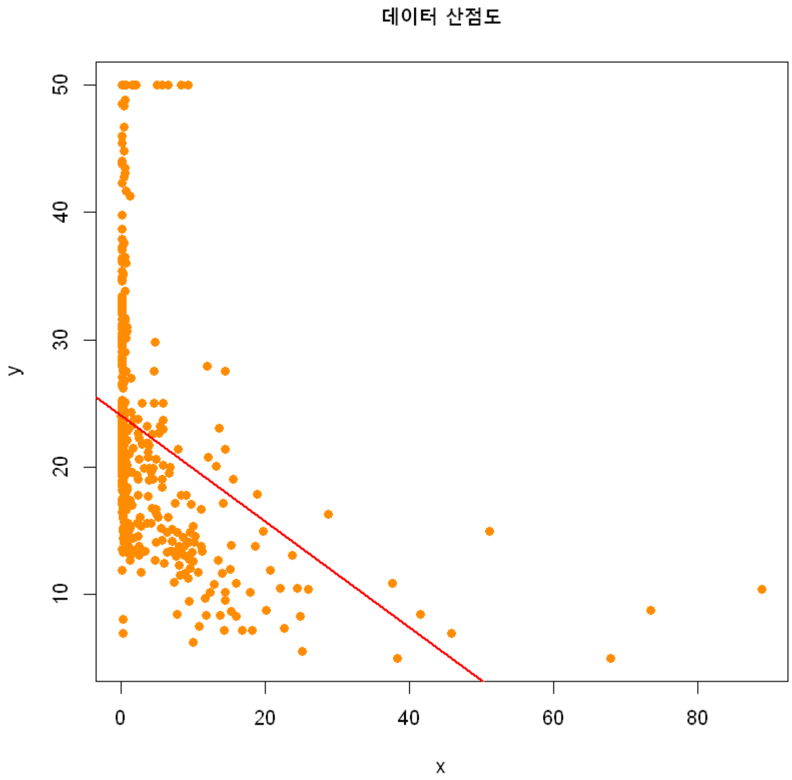
```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-16.957  -5.449  -2.007   2.512  29.800 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 24.03311    0.40914   58.74   <2e-16 ***
x           -0.41519    0.04389   -9.46   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.484 on 504 degrees of freedom
Multiple R-squared:  0.1508,	Adjusted R-squared:  0.1491 
F-statistic: 89.49 on 1 and 504 DF,  p-value: < 2.2e-16
```

In [5]:
```
plot(y~x, pch = 16, col = 'darkorange', main = '데이터 산점도')
abline(model, lwd = 2, col = 'red')
```

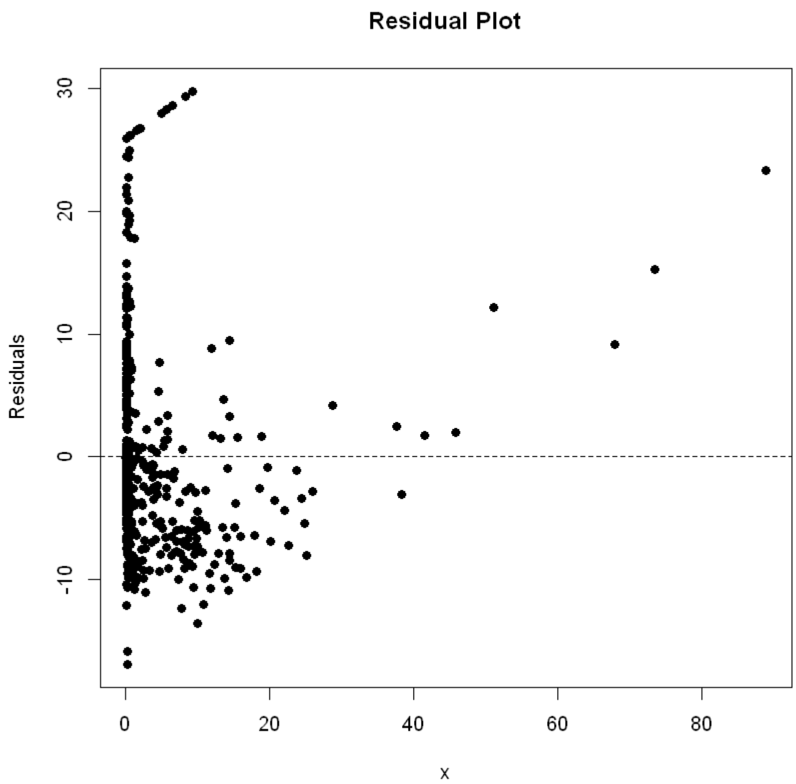


> `p-value` 로 보았을 때 모형은 통계적으로 유의하나, 잘 적합된 것 같아 보이진 않는다.

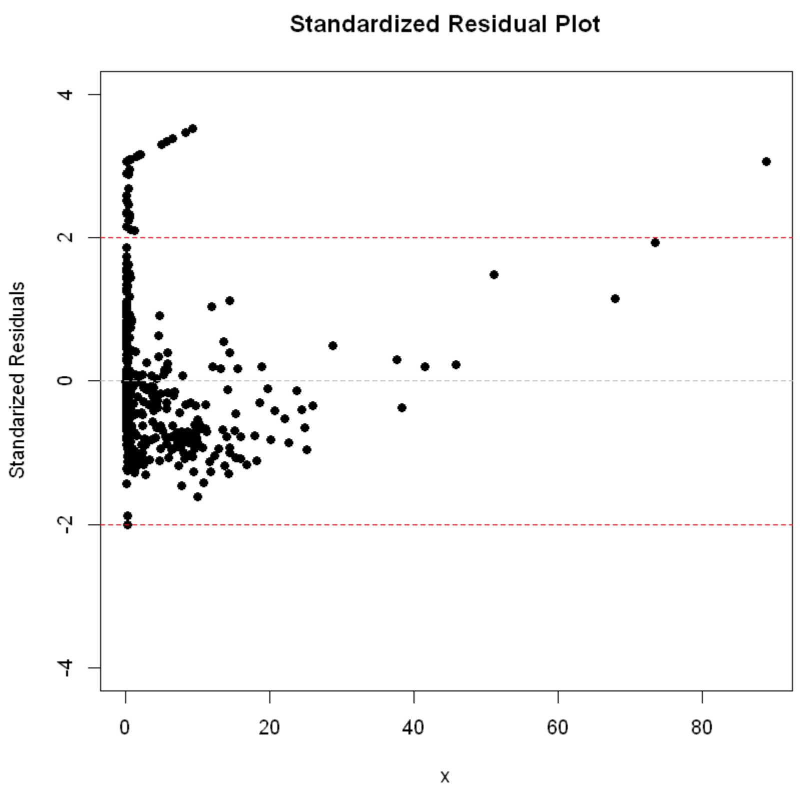
---

(2) 잔차의 산점도를 그려보고 모형의 타당성과 등분산성에 대하여 설명하시오.

In [6]:
```
plot(model$residuals ~ x, pch = 16, ylab = 'Residuals', main = 'Residual Plot')
abline(h = 0, lty = 2)
```



In [7]:
```
plot(rstandard(model) ~ x, pch = 16, ylab = 'Standarized Residuals', ylim = c(-4
abline(h = c(-2, 0, 2), col = c('red', 'grey', 'red'), lty = 2)
```



i) 선형성 : 모형이 0을 중심으로 대칭인 것처럼 보이진 않는다.

ii) 등분산성 : 설명변수 x값에 관계없이 잔차의 분산이 일정해야 하는데, 이 경우 x값이 작을수록 잔차의 분산이 커지고 있다. 따라서 모형의 등분산성 가정은 옳지 못하다.

iii) 정규성 : 표준화된 잔차를 산점도로 나타냈을 때 2를 넘는 값이 많이 관측되며, 3을 넘는 값들도 꽤 많이 보인다. 이에 따라 정규성을 가정하기 어려울 것으로 예상된다.

iv) 독립성 : 잔차가 x의 값에 따라 비선형의 패턴이 있어보인다. 이에 따라 독립성을 가정하기 어려울 것으로 예상된다.

> 따라서 선형회귀의 기본가정에 위배되므로, 단순선형회귀로 적합한 모형은 타당하지 않은 것 같다.

---

(3) 유의수준 $\alpha = 0.1$에서 $H_0 : \beta_1 = 0.3$, $H_1 : \beta_1 \neq 0.3$을 검정하시오

In [8]:
```
## T-통계량
t_0 <- (as.numeric(model$coefficients[2]) - 0.3)/summary(model)$coefficients[2,2
print(paste('T-statiscic =', t_0))

## p-value
p_value <- pt(q = t_0, df = length(x-2)) + pt(q = -t_0, df = length(x-2), lower.
print(paste('p_value =', p_value))

## 검정
print(paste('Result of \'p_value < 0.1\' :', p_value < 0.1,'// So reject null hy
```

```
[1] "T-statiscic = -16.2949207615693"
[1] "p_value = 2.69319568160759e-48"
[1] "Result of 'p_value < 0.1' : TRUE // So reject null hypothesis"
```

> 유의수준 $\alpha = 0.1$에서 `p_value` 가 α보다 작으므로, 귀무가설을 기각하고, 대립가설을 수용한다. 즉, β_1은 0.3이 아니다.

---

(4) $Durbin - Watson\ \ d$통계량을 사용하여 $H_0 : \rho = 0, H_0 : \rho > 0$을 유의수준 $\alpha = 0.05$에서 검정하시오

In [10]:
```
dwtest(model, alternative = 'greater')
```

```
	Durbin-Watson test

data:  model
DW = 0.71342, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

> `p-value` 가 $\alpha = 0.05$보다 작기 때문에 귀무가설을 기각한다. 따라서 잔차는 1차 양의 자기상관을 지닌다.