Part 2. R 통계분석 (데이터 분석 전문가 양성과정)

13

선형회귀의 이해

경북대학교 배준현 교수
(joonion@knu.ac.kr)

13. 선형회귀의 이해

- **회귀**: *regression*
  - '회귀'의 사전적 의미: 되돌아감(어디로?)
  - 회귀라는 용어의 유래:
    - 프랜시스 골턴의 유전학 연구에서 유래함
    - 회귀의 법칙: *the law of regression*
  - 프랜시스 골턴의 연구:
    - 부모의 키와 자녀의 키는 유전적으로 어떤 관계가 있는가?
    - 평균으로의 회귀: *regression to the mean*

13. 선형회귀의 이해

TABLE I.

NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES.

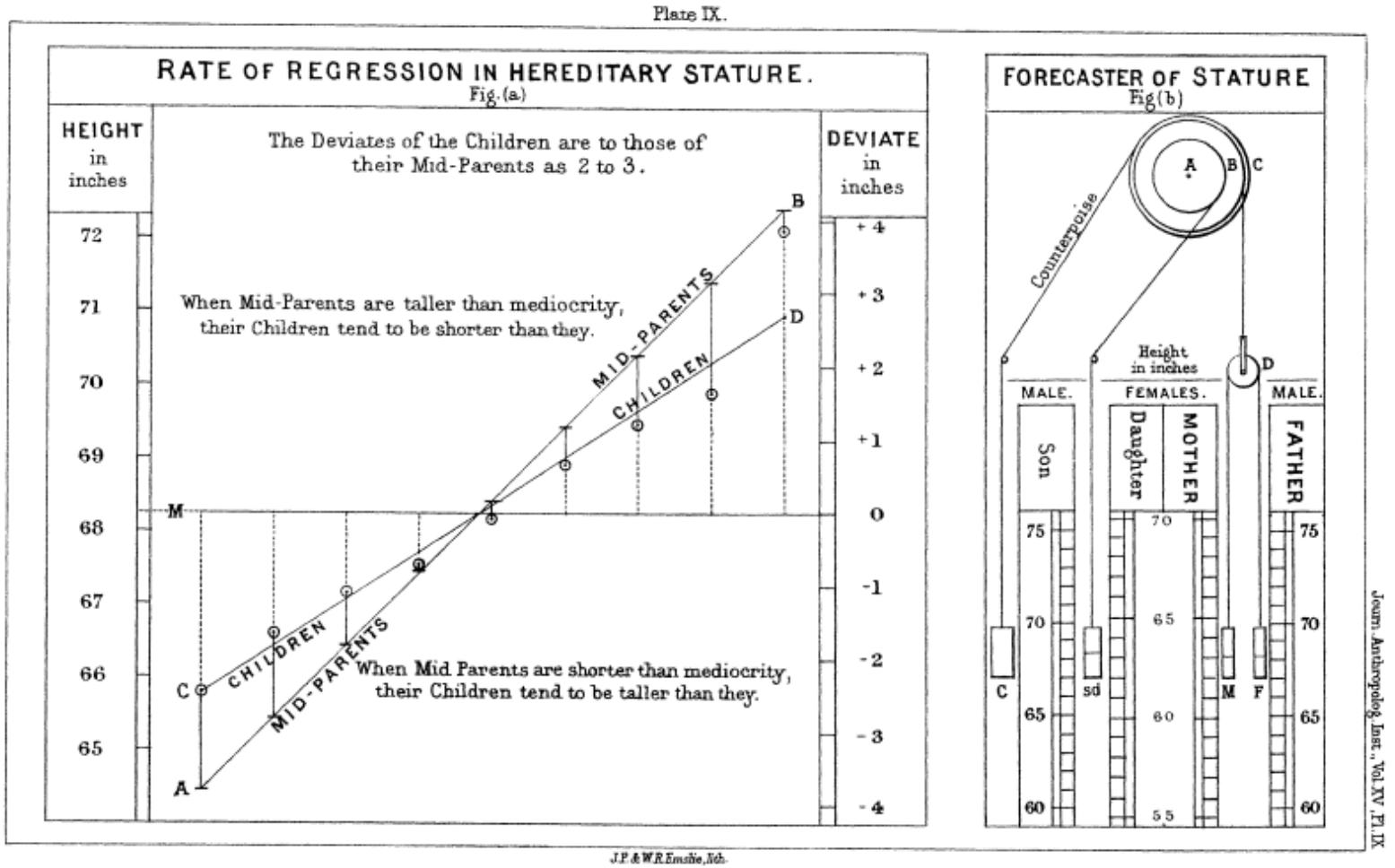
(All Female heights have been multiplied by 1·08).

Heights of the Mid-parents in inches.	Heights of the Adult Children.														Total Number of		Medians.
	Below	62·2	63·2	64·2	65·2	66·2	67·2	68·2	69·2	70·2	71·2	72·2	73·2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72·5	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72·2
71·5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69·9
70·5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69·5
69·5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68·9
68·5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68·2
67·5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67·6
66·5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67·2
65·5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66·7
64·5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65·8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66·3	67·8	67·9	67·7	67·9	68·3	68·5	69·0	69·0	70·0	..	..	..	..	..

NOTE.—In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62·2, 63·2, &c., instead of 62·5, 63·5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents.

Galton, Francis. "Regression towards mediocrity in hereditary stature." *The Journal of the Anthropological Institute of Great Britain and Ireland* 15 (1886): 246-263.

13. 선형회귀의 이해



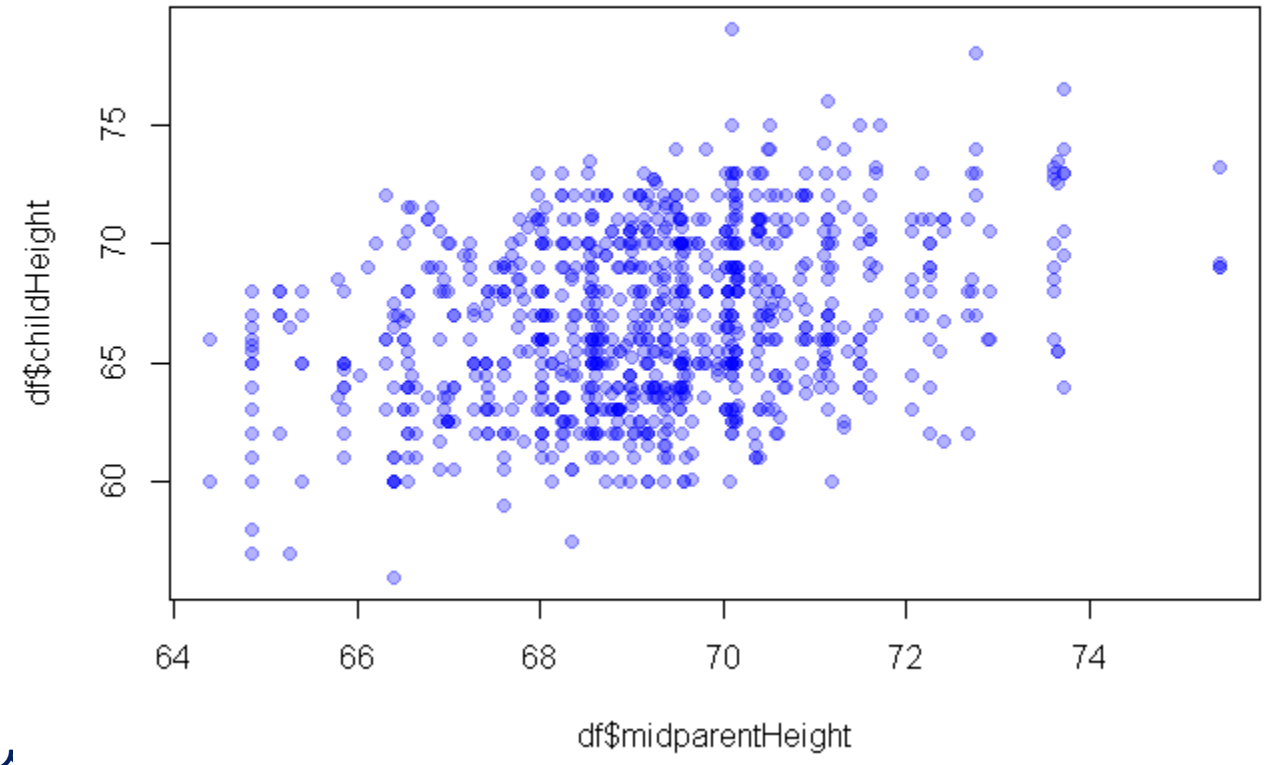
Galton, Francis. "Regression towards mediocrity in hereditary stature." *The Journal of the Anthropological Institute of Great Britain and Ireland* 15 (1886): 246-263.

프랜시스 골턴의 데이터셋: *GaltonFamilies*

```
> library(HistData)
> str(GaltonFamilies)
'data.frame':  934 obs. of  8 variables:
 $ family         : Factor w/ 205 levels "001","002","003",..: 1 1 1 1 2 2 2 2 3 3 ...
 $ father         : num  78.5 78.5 78.5 78.5 75.5 75.5 75.5 75.5 75 75 ...
 $ mother         : num  67 67 67 67 66.5 66.5 66.5 66.5 64 64 ...
 $ midparentHeight: num  75.4 75.4 75.4 75.4 73.7 ...
 $ children       : int  4 4 4 4 4 4 4 4 2 2 ...
 $ childNum       : int  1 2 3 4 1 2 3 4 1 2 ...
 $ gender         : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 1 1 2 1 ...
 $ childHeight    : num  73.2 69.2 69 69 73.5 72.5 65.5 65.5 71 68 ...
```

13. 선형회귀의 이해

```
> df <- GaltonFamilies
> plot(df$midparentHeight, df$childHeight,
       pch = 19, col = adjustcolor("blue", alpha.f = 0.3))
```

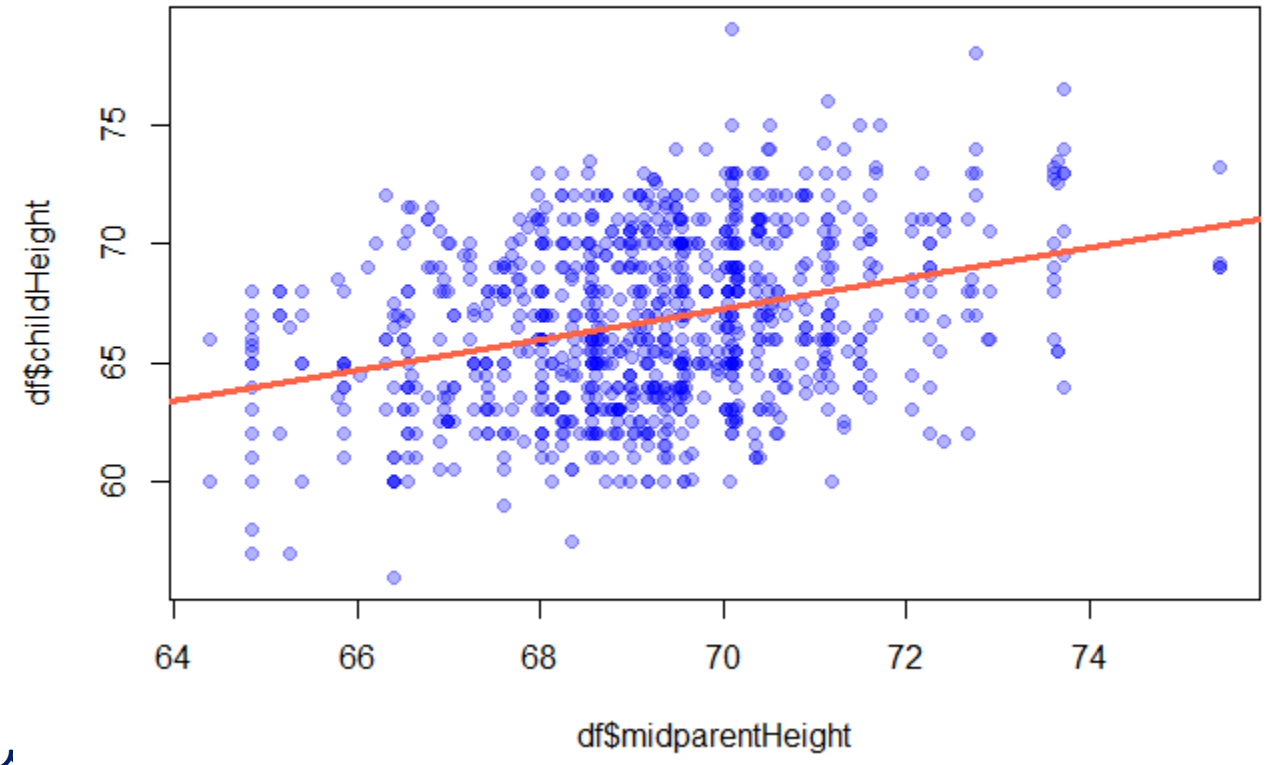


13. 선형회귀의 이해

```
> cor(df$midparentHeight, df$childHeight)
[1] 0.3209499

> model <- lm(childHeight ~ midparentHeight, data = df)
> model

Call:
lm(formula = childHeight ~ midparentHeight, data = df)

Coefficients:
      (Intercept)  midparentHeight
          22.6362           0.6374
```

13. 선형회귀의 이해

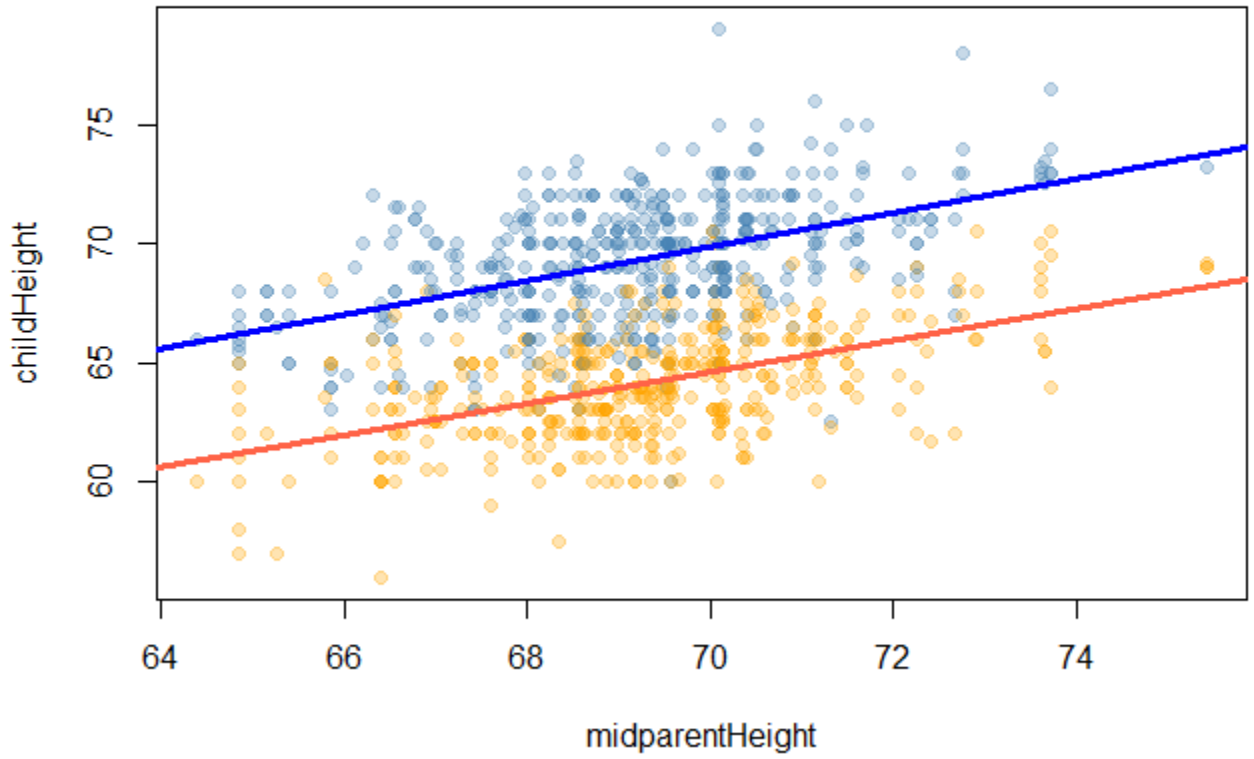
```
> plot(df$midparentHeight, df$childHeight,
       pch = 19, col = adjustcolor("blue", alpha.f = 0.3))
> abline(model, col = "tomato", lty = 1, lwd = 3)
```



- 자녀의 성별에 따라 키의 분포도 달라지지 않을까?

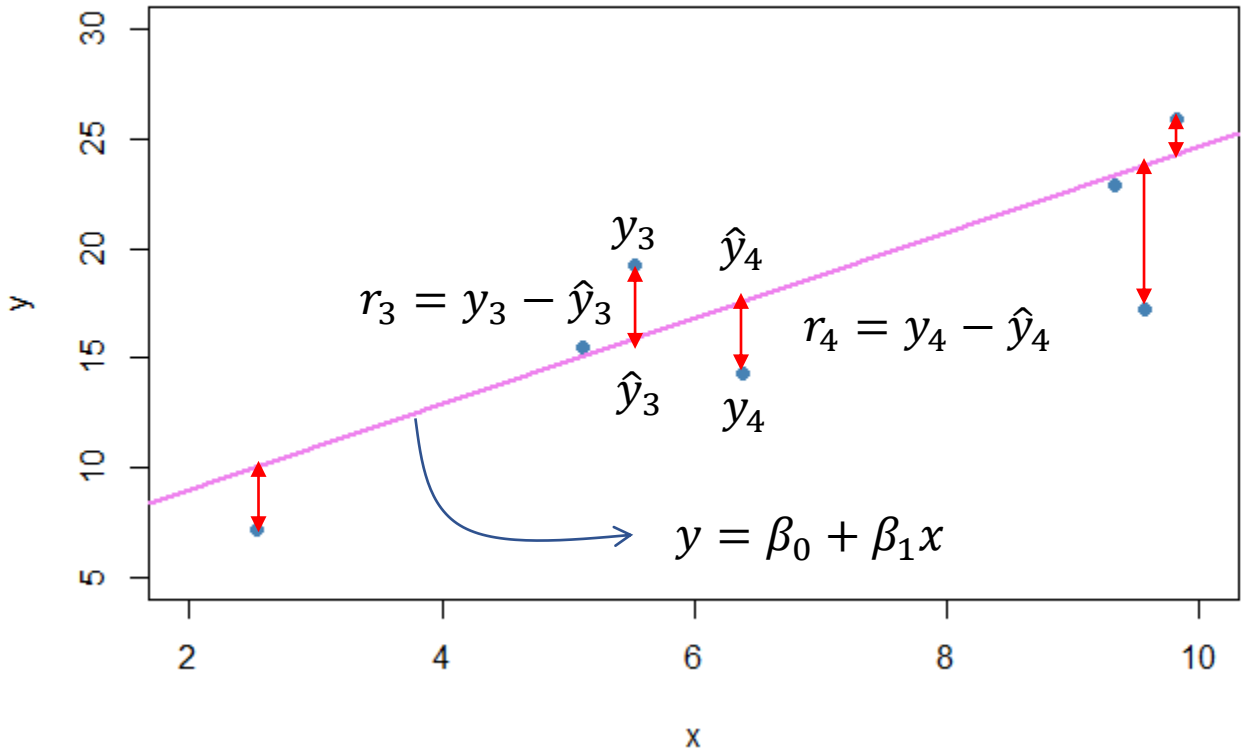
```
> color.m <- adjustcolor("steelblue", alpha.f = 0.3)
> color.f <- adjustcolor("orange", alpha.f = 0.3)

> with(df,
       plot(midparentHeight, childHeight, pch = 19,
            col = ifelse(gender == "male", color.m, color.f)))

> model.m <- lm(childHeight ~ midparentHeight,
                data = subset(df, gender == "male"))
> abline(model.m, col = "blue", lty = 1, lwd = 3)

> model.f <- lm(childHeight ~ midparentHeight,
                data = subset(df, gender == "female"))
> abline(model.f, col = "tomato", lty = 1, lwd = 3)
```

13. 선형회귀의 이해



- **회귀분석과 선형회귀:**
  - 회귀분석: *regression analysis*
    - 독립변수와 종속변수의 관계를 잘 설명하는 회귀식을 찾는 과정
  - 선형회귀: *linear regression*
    - 독립변수와 종속변수의 관계가 선형일 때
    - 선형 회귀식(직선의 방정식): $y = \beta + \alpha x$
    - 선형 회귀식의 절편(*intercept*)과 기울기(*slope*)를 알면
      - 독립변수와 종속변수의 관계를 설명, 또는, 예측할 수 있다.

- **선형 회귀모델:** *linear regression model*
  - **회귀식:** $y = \beta_0 + \beta_1 x$
  - **잔차**(*residual*)**: 실제 데이터의 값**(관측값)**과 회귀식의 값**(예측값)**과의 차이**
    - $r_i = y_i - \hat{y}_i$, r_i: **잔차**, y_i: **관측값**, $\hat{y}_i$: **예측값**



```
> set.seed(14)
> x <- runif(n = 7, min = 0, max = 10)
> y <- 3 + 2 * x + rnorm(n = 7, mean = 0, sd = 5)
> round(x, 2)
[1] 2.54 6.38 9.57 5.53 9.83 5.11 9.33
> round(y, 2)
[1]  7.18 14.25 17.25 19.26 25.87 15.48 22.86
```

i	1	2	3	4	5	6	7
x_i	2.54	6.38	9.57	5.53	9.83	5.11	9.33
y_i	7.18	14.25	17.25	19.26	25.87	15.48	22.86
$\hat{y}_i$							
r_i							

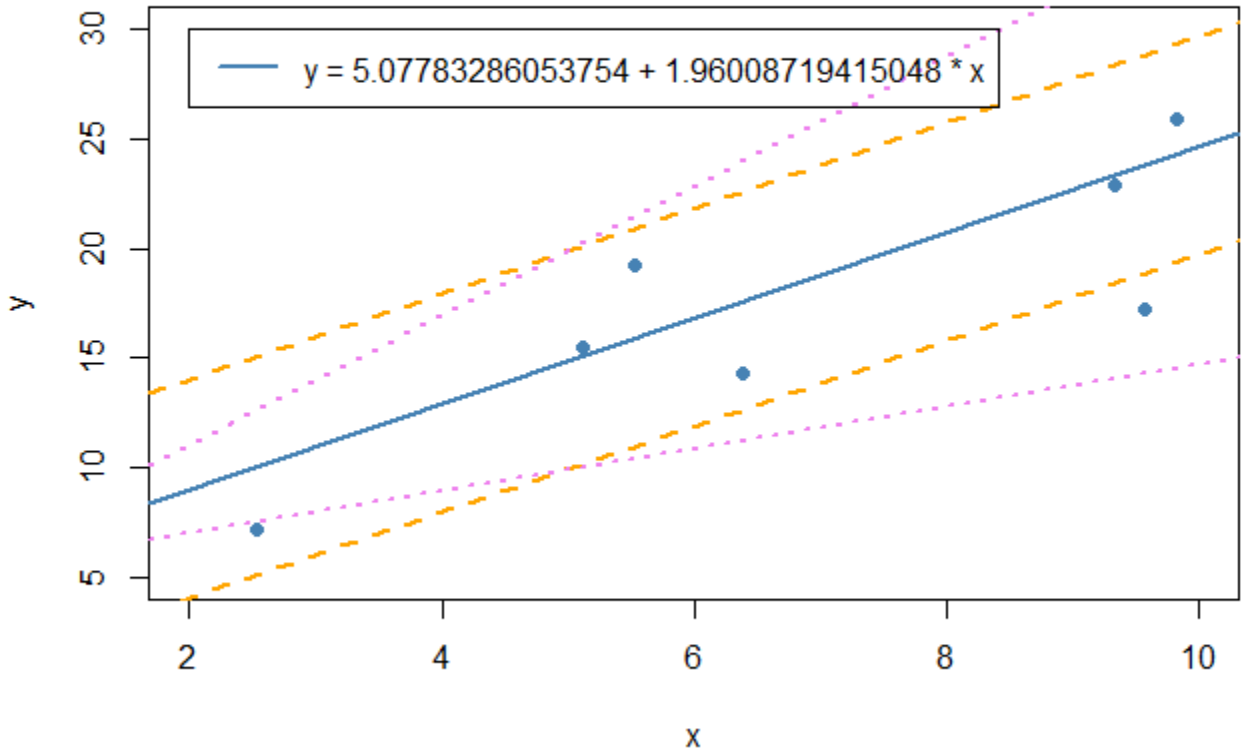
```
> model <- lm(y ~ x, data = df)
> coef(model)
(Intercept)           x
   5.077833    1.960087
> intercept <- coef(model)[1]
> slope <- coef(model)[2]
> y.hat <- intercept + slope * x
> round(y.hat, 2)
[1] 10.06 17.58 23.84 15.91 24.35 15.10 23.36
> r <- y - y.hat
> round(r, 2)
[1] -2.88 -3.33 -6.59  3.35  1.53  0.37 -0.50
```

i	1	2	3	4	5	6	7
x_i	2.54	6.38	9.57	5.53	9.83	5.11	9.33
y_i	7.18	14.25	17.25	19.26	25.87	15.48	22.86
$\hat{y}_i$	10.06	17.58	23.84	15.91	24.35	15.10	23.36
r_i	-2.88	-3.33	-6.59	3.35	1.53	0.37	-0.50

- **모형 적합**: *fitting* a model
  - 데이터(관측값)를 가장 잘 설명하는 선형 회귀식은?
    - 데이터 전체를 고려했을 때 잔차가 가장 작은 직선의 방정식
  - **평균절대오차**: MAE, mean absolute error
    - $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$
  - **평균제곱오차**: MSE, mean squared error
    - $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
  - **제곱근 평균제곱오차**: RMSE, rooted mean squared error
    - $RMSE = \sqrt{MSE}$

```
> plot(x, y, pch = 19, col = "steelblue", xlim = c(2, 10), ylim = c(5, 30))
> abline(model, lwd = 2, col = "steelblue")
> abline(a = intercept + 5, b = slope, lty = 2, lwd = 2, col = "orange")
> abline(a = intercept - 5, b = slope, lty = 2, lwd = 2, col = "orange")
> abline(a = intercept, b = slope + 1, lty = 3, lwd = 2, col = "violet")
> abline(a = intercept, b = slope - 1, lty = 3, lwd = 2, col = "violet")
> legend(x = 2, y = 30, lwd = 2, col = "steelblue",
         legend = paste("y =", intercept, "+", slope, "* x"))
```

13. 선형회귀의 이해


y = 5.07783286053754 + 1.96008719415048 * x
y
x
5
10
15
20
25
30
2
4
6
8
10

Any Questions?