데이터 과학 기초

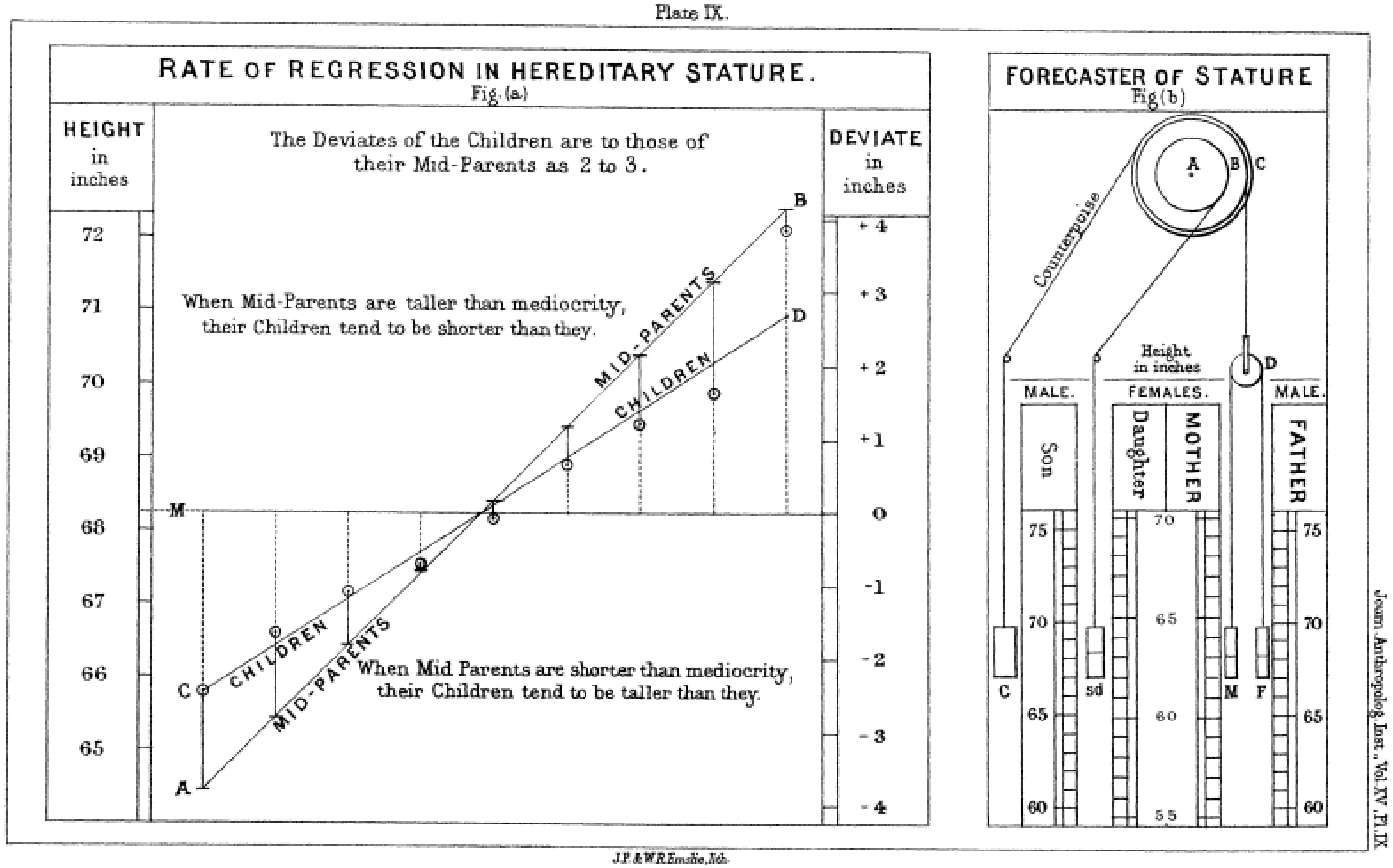
03

선형 회귀

경북대학교 배준현 교수
(joonion@knu.ac.kr)

- **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*

03. 선형 회귀



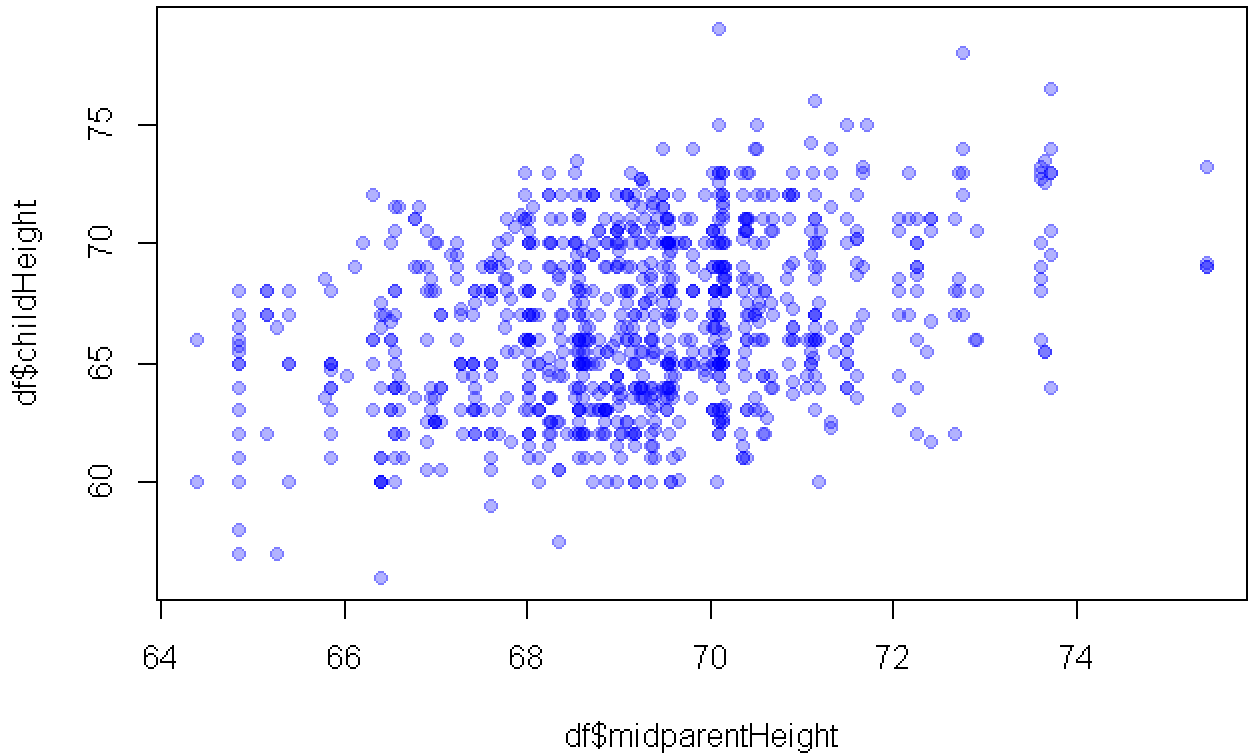
Galton, Francis. "Regression towards mediocrity in hereditary stature." *The Journal of the Anthropological Institute of Great Britain and Ireland* 15 (1886): 246-263.

- **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 ...
```

03. 선형 회귀

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

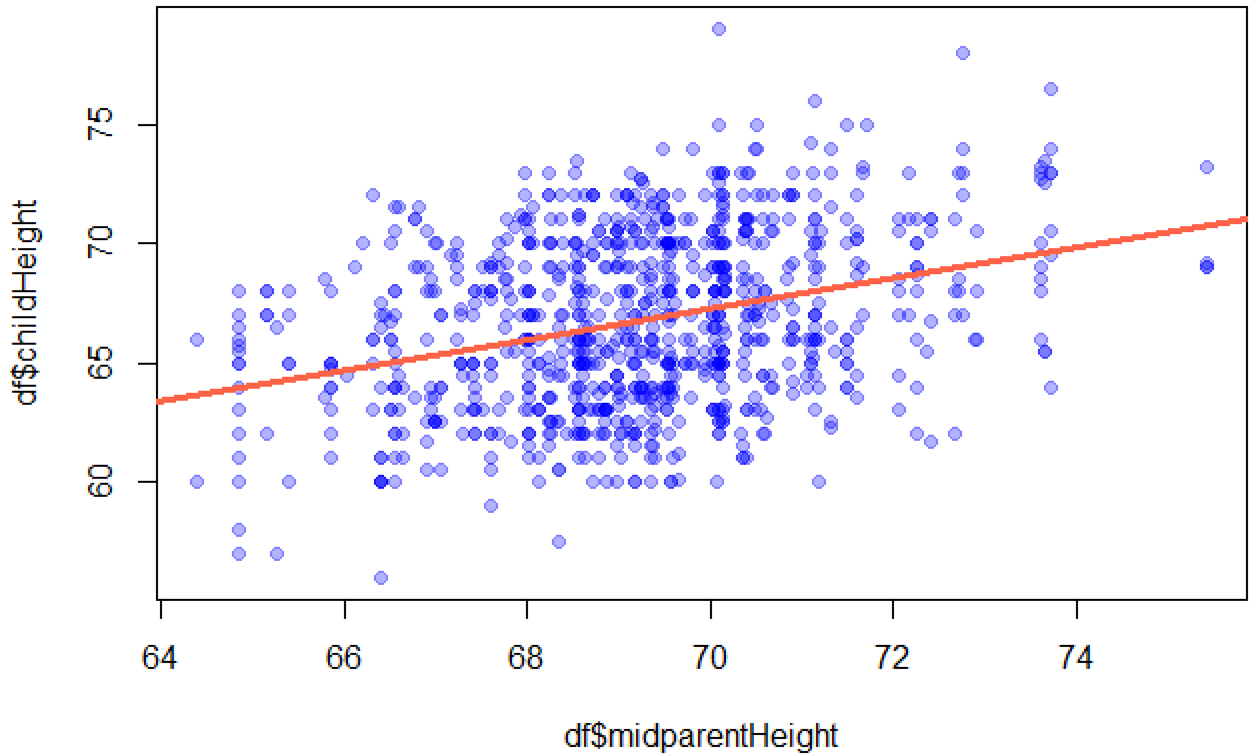


03. 선형 회귀

```
> 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
```

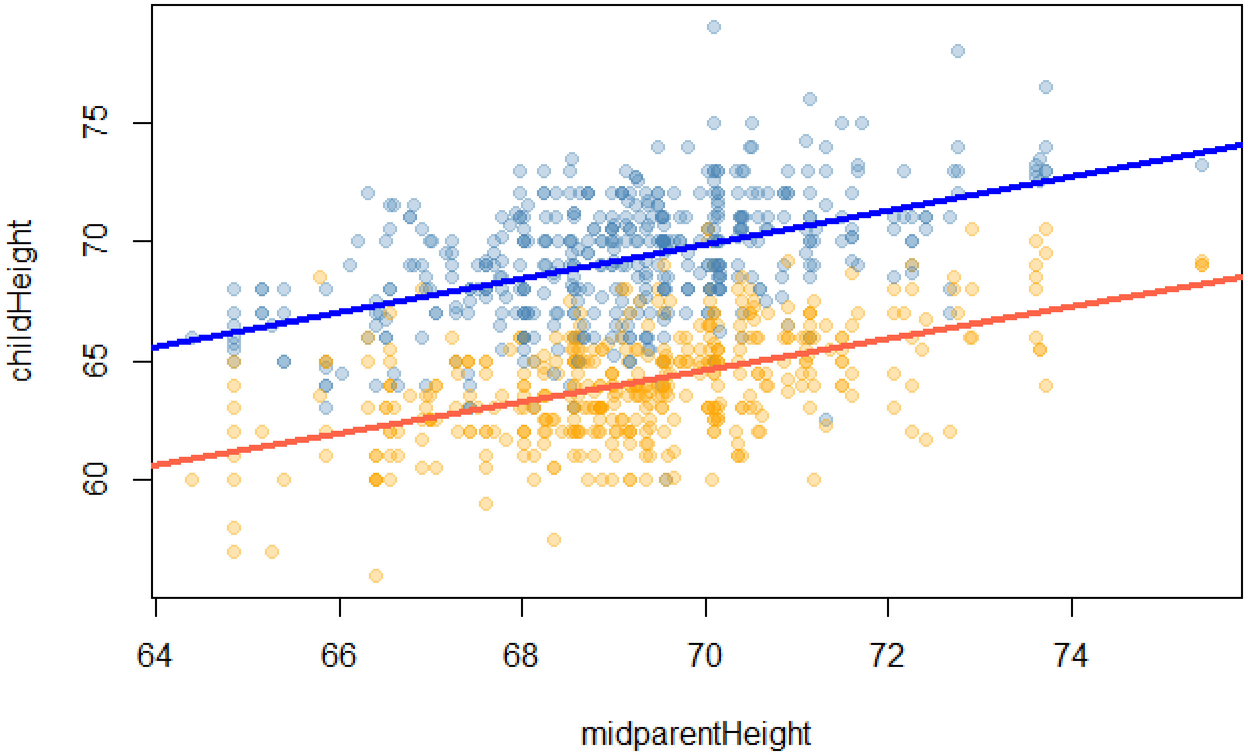
03. 선형 회귀

```
> 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?

```r
> 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)
```



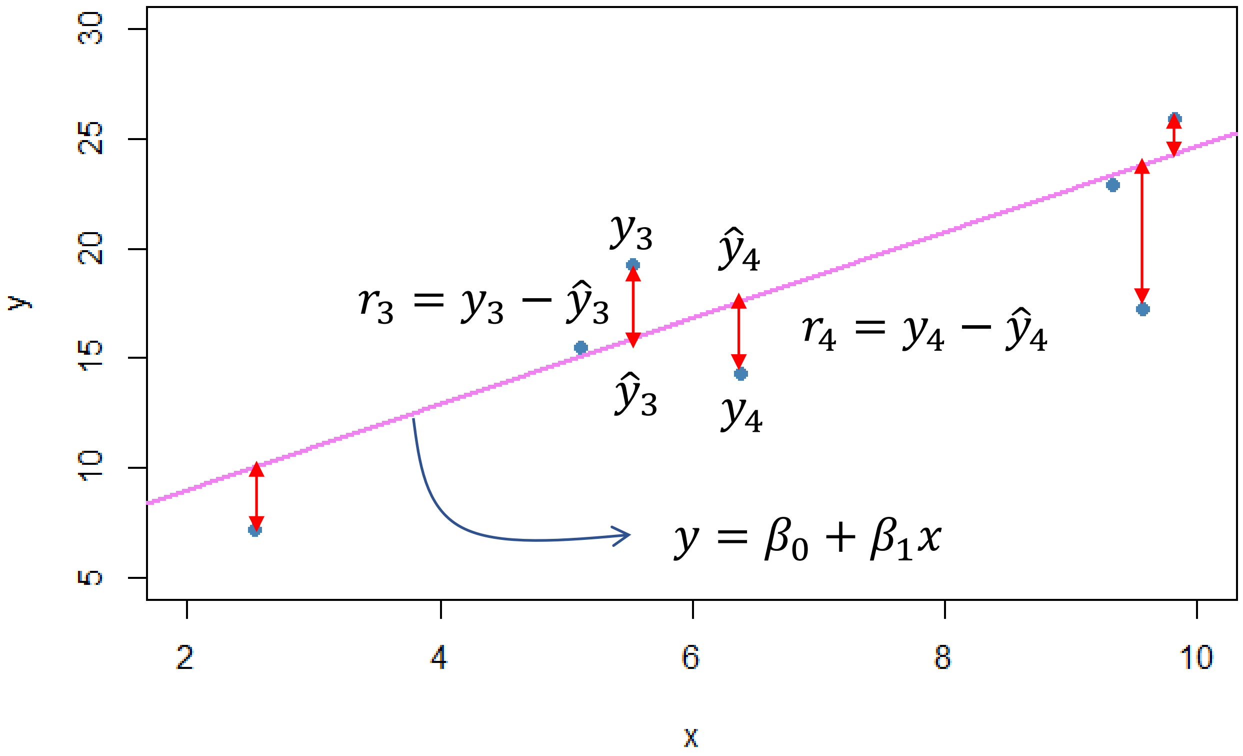
childHeight
60
65
70
75
64
66
68
70
72
74
midparentHeight

03. 선형 회귀

- **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



03. 선형 회귀


```
> 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							

03. 선형 회귀


```
> 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
```

03. 선형 회귀

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

03. 선형 회귀

- **선형회귀의 유형:**
  - 단순 선형회귀: *simple(univariate) linear regression*
    - 한 개의 독립변수와 종속변수 간의 단순한(일차) 선형 관계
  - 다중 선형회귀: *multiple(multivariate)* linear regression
    - 두 개 이상의 독립변수와 종속변수 간의 선형 관계
  - 다항 선형회귀: *polynomial linear regression*
    - 종속변수와 한 개의 독립변수의 다항식으로 구성된 비선형 관계

- 단순 선형회귀: *simple* linear regression
  - 교육기간과 평균소득 간에는 선형 관계가 있을까?
    - 종속변수: 평균소득(income)
    - 독립변수: 교육기간(education)

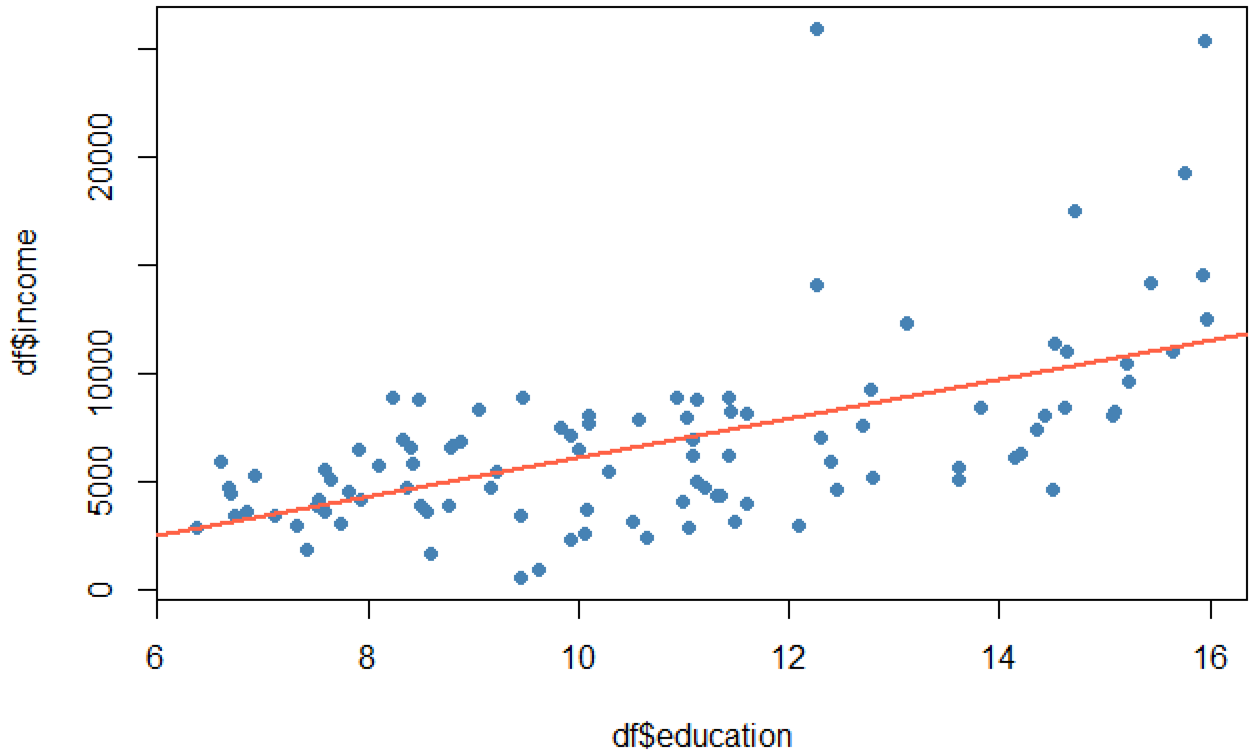
```
> library(car)
> str(Prestige)
'data.frame':	102 obs. of  6 variables:
 $ education: num  13.1 12.3 12.8 11.4 14.6 ...
 $ income   : int  12351 25879 9271 8865 8403 11030 8258 14163 11377 11023 ...
 $ women    : num  11.16 4.02 15.7 9.11 11.68 ...
 $ prestige : num  68.8 69.1 63.4 56.8 73.5 77.6 72.6 78.1 73.1 68.8 ...
 $ census   : int  1113 1130 1171 1175 2111 2113 2133 2141 2143 2153 ...
 $ type     : Factor w/ 3 levels "bc","prof","wc": 2 2 2 2 2 2 2 2 2 2 ...
```

03. 선형 회귀

- Prestige 데이터셋
  - 캐나다의 인구조사 데이터(1971년): 변수 6개, 관측값 102개
    - education: 재직자의 평균 교육기간 (years)
    - income: 재직자의 평균 소득 (dollars)
    - women: 여성 재직자의 비율
    - prestige: 직업에 대한 명성 점수 (1960년대 중반에 실시된 사회 조사 결과)
    - census: 캐나다의 직업 코드
    - type: 직업 분류: bc: blue color, prof: professional, wc: white color

03. 선형 회귀

```
> model <- lm(formula = formula, data = Prestige)
> abline(model, lwd = 2, col = "tomato")
```



- **다중 선형회귀**: *multiple* linear regression
  - 종속변수에 영향을 미치는 독립변수가 여러 개일 경우
    - 다중 회귀식: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$
  - 평균소득에 영향을 주는 요인은 무엇일까?
    - 종속변수: 평균소득(income)
    - 독립변수: 교육(education), 성별(women), 명성(prestige)

income ~ education + women + prestige

03. 선형 회귀

```
> model <- lm(income ~ ., data = df)
> summary(model)

Call:
lm(formula = income ~ ., data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-7715.3  -929.7  -231.2   689.7 14391.8 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -253.850   1086.157  -0.234    0.816    
education    177.199    187.632   0.944    0.347    
women        -50.896      8.556  -5.948 4.19e-08 ***
prestige     141.435     29.910   4.729 7.58e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2575 on 98 degrees of freedom
Multiple R-squared:  0.6432,	Adjusted R-squared:  0.6323 
F-statistic: 58.89 on 3 and 98 DF,  p-value: < 2.2e-16
```

03. 선형 회귀

- 다항 선형회귀: *polynomial* linear regression
  - 종속변수를 독립변수의 다항식이 더 잘 설명하는 경우
    - 다항 회귀식: $y = \beta_0 + \beta_1 x^1 + \beta_2 x^2 + \cdots + \beta_n x^n$
  - 교육기간과 평균소득의 관계를 직선보다 더 잘 설명하는 곡선이 있을까?
    - 종속변수: 평균소득(income)
    - 독립변수: 교육기간(education)

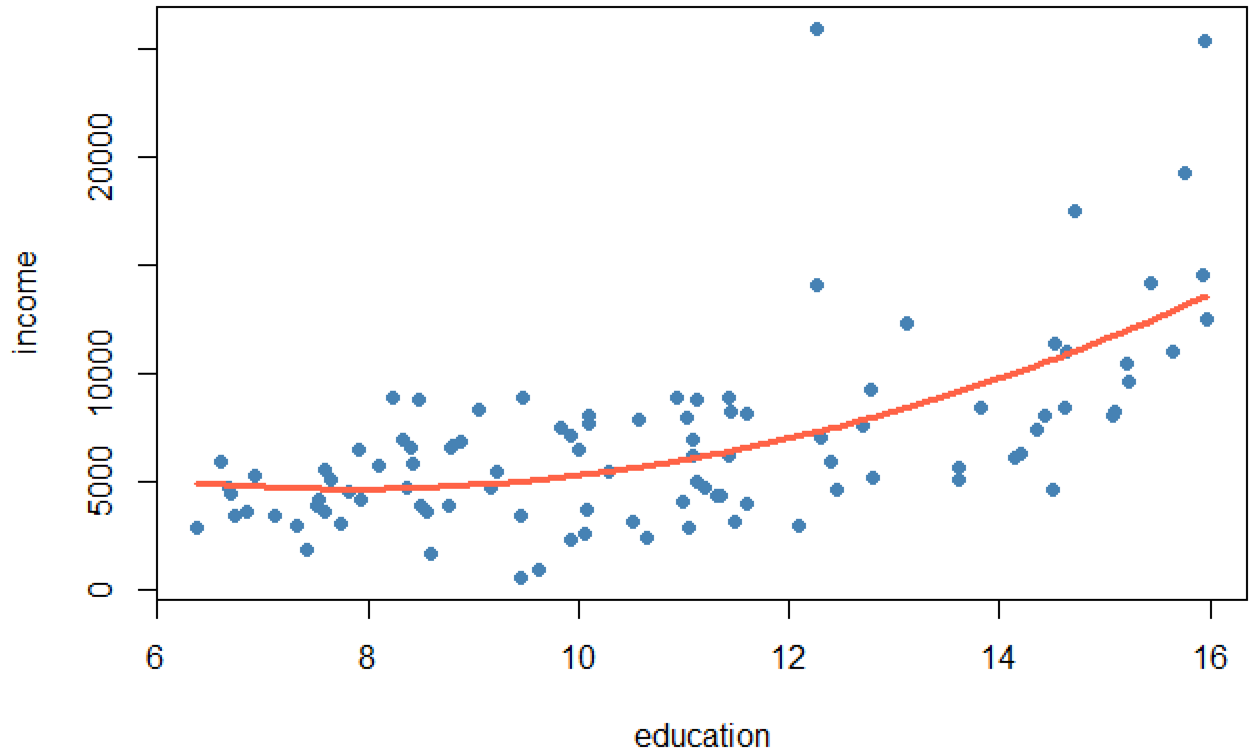
- 다항 선형회귀: *polynomial* linear regression
  - 종속변수를 독립변수의 다항식이 더 잘 설명하는 경우
    - 다항 회귀식: $y = \beta_0 + \beta_1 x^1 + \beta_2 x^2 + \cdots + \beta_n x^n$
  - 교육기간과 평균소득의 관계를 직선보다 더 잘 설명하는 곡선이 있을까?
    - 종속변수: 평균소득(income)
    - 독립변수: 교육기간(education)

03. 선형 회귀

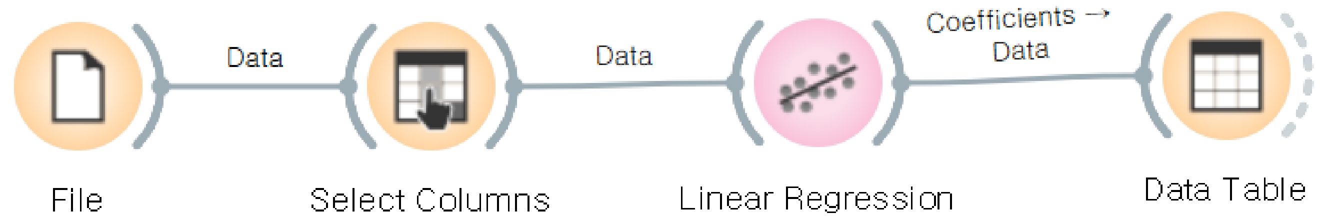
```
> library(car)
> formula <- income ~ education + I(education^2)
> model <- lm(formula, data = Prestige)
> summary(model)

Call:
lm(formula = formula, data = Prestige)

Residuals:
    Min      1Q  Median      3Q     Max 
-5951.4 -2091.1  -358.2  1762.4 18574.2 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)   
(Intercept)    12918.23    5762.27   2.242  0.02720 * 
education      -2102.90    1072.73  -1.960  0.05277 . 
I(education^2)   134.18      47.64   2.817  0.00586 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3369 on 99 degrees of freedom
Multiple R-squared:  0.383,	Adjusted R-squared:  0.3706 
F-statistic: 30.73 on 2 and 99 DF,  p-value: 4.146e-11
```

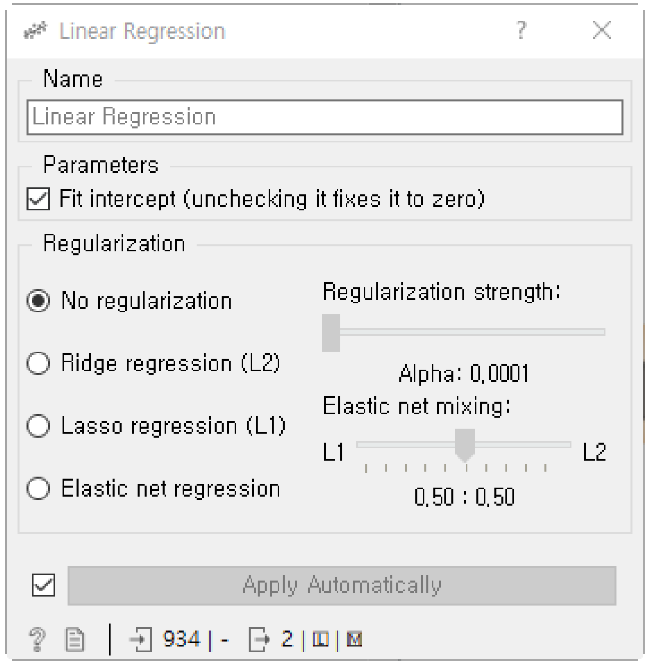
03. 선형 회귀

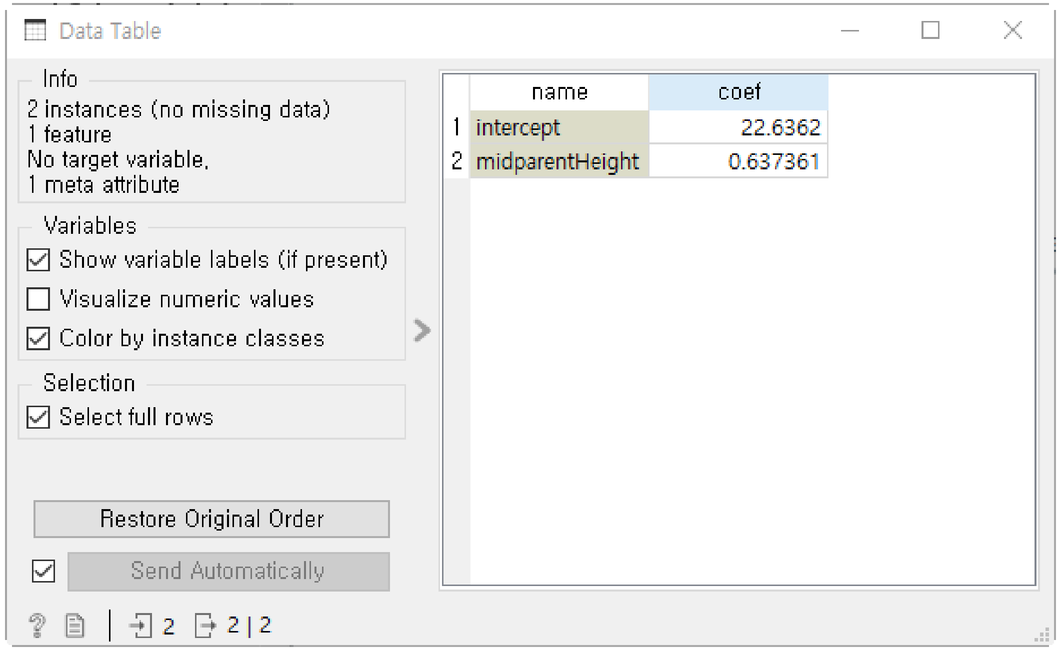
```
> plot(income ~ education, data = Prestige, pch = 19, col = "steelblue")
> library(dplyr)
> with(Prestige,
       lines(arrange(data.frame(education, fitted(model)), education),
             lty = 1, lwd = 3, col = "tomato"))
```



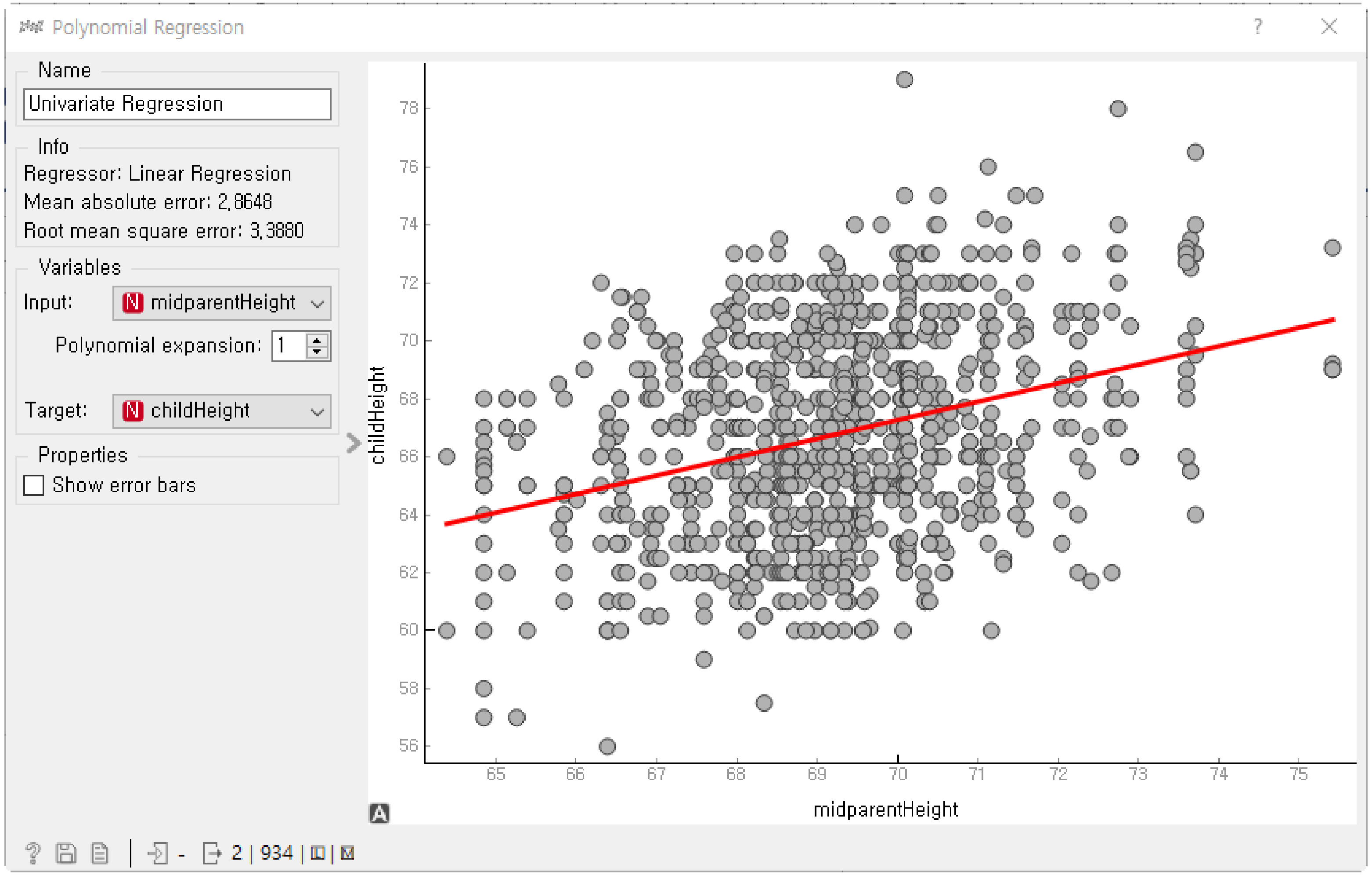
Orange: Linear Regression



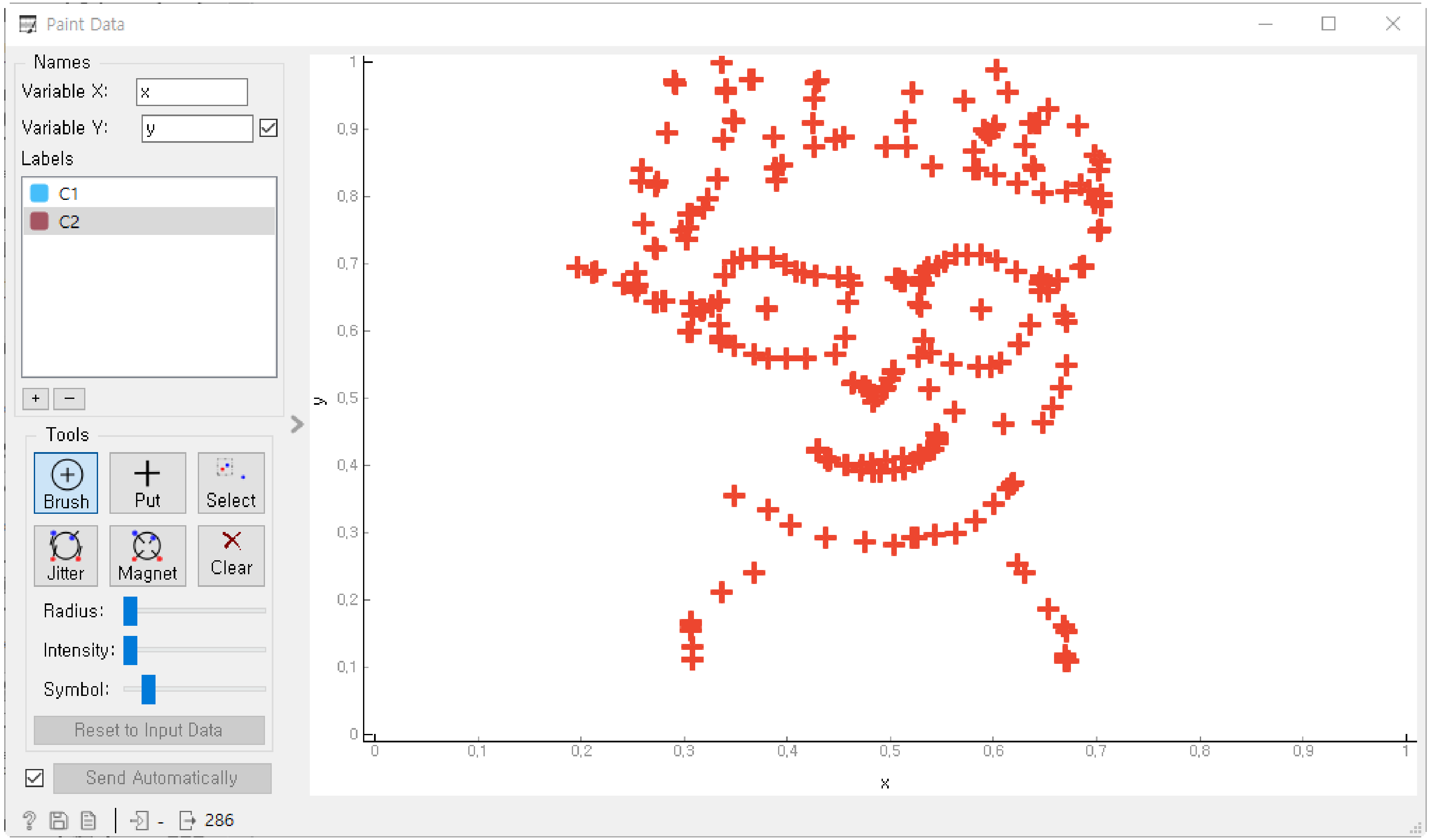




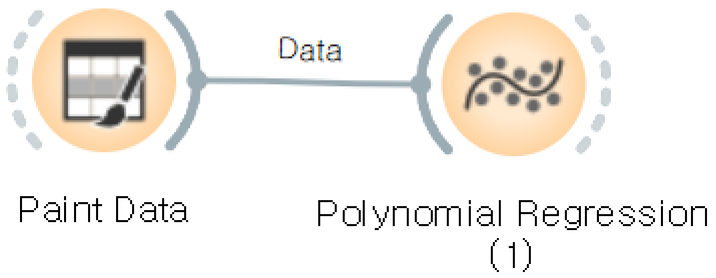
- Orange: Educational/Polynomial Regression

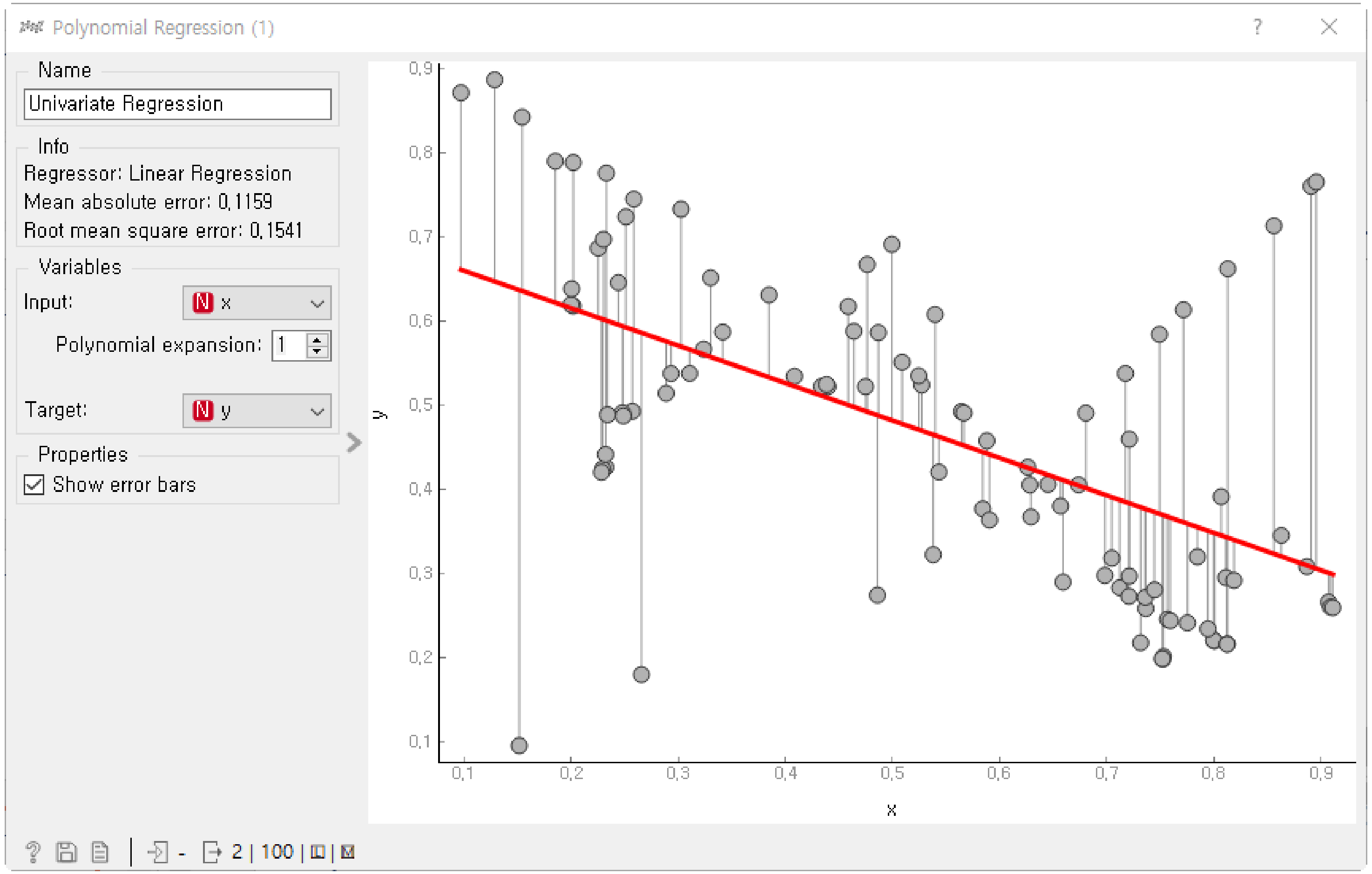


- Orange: Paint Data



▪ Orange: 다양한 선형 회귀 실험





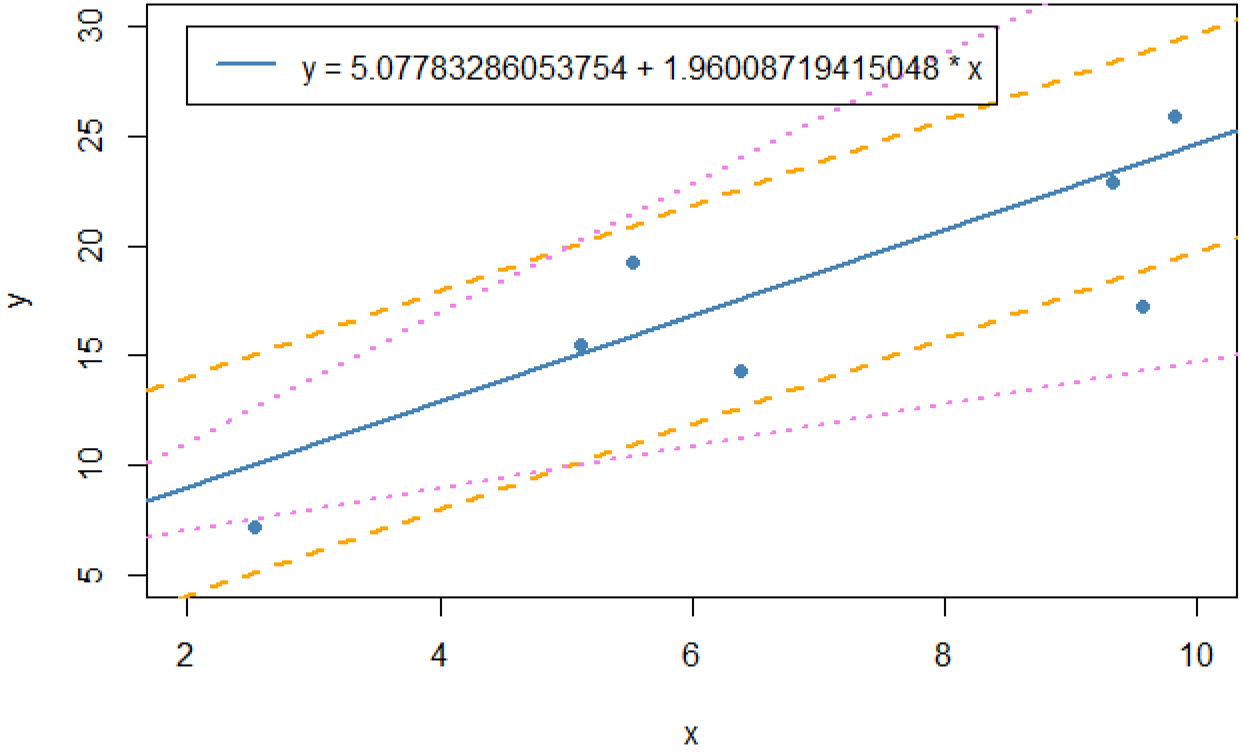
03. 선형 회귀

- **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
    - $RMSE = \sqrt{MSE}$

03. 선형 회귀

```r
> 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"))
```

03. 선형 회귀



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

03. 선형 회귀

- **결정계수**: *coefficient of determination*
  - R^2(*R-squared*): **선형 회귀식의 설명력 지표**
    - $R^2 = \frac{SSE(Explained\ Sum\ of\ Squares)}{SST(Total\ Sum\ of\ Squares)} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$
    - $R^2 = 0$: **독립변수와 종속변수 간의 선형 관계가 존재하지 않음**
    - $R^2 = 1$: **독립변수와 종속변수 간에는 완전한 선형 관계가 존재함**
  - *Adjusted* R^2: **다중 독립변수의 영향을 줄여줌**
    - R^2**는 독립변수의 개수가 증가하면 항상 값이 증가함**
    - **독립변수의 개수가 많아지면 페널티를 부과하여 설명력을 보정**
    - **과적합**(*overfitting*)**에 대한 고려**

- 선형회귀 모델을 적용하기 위한 전제 조건:
  - 선형성: *linearity*
    - 독립변수와 종속변수 간의 선형적 관계가 존재한다.
  - 정규성: *normality*
    - 종속변수의 값들이 정규분포를 가진다.
  - 등분산성: *homoscedasticity*, *homogeneity of variance*
    - 종속변수 값들의 분포는 모두 동일한 분산을 가진다.
  - 독립성: *independence*
    - 모든 독립변수의 관측값들은 서로 독립이다.

03. 선형 회귀

- **페널티 회귀분석**: *penalized* regression analysis
  - 너무 많은 독립변수를 갖는 모델에 페널티를 부과하여 간명한 회귀모델을 생성
    - 모델의 성능에 크게 기여하지 못하는 변수의 영향력을 축소하거나 제거
    - 제약화, 규제화: regularization, 축소: shirinkage
  - 회귀식에 페널티항을 추가
    - 잔차제곱합과 페널티항의 합이 최소가 되는 회귀계수를 추정
  - 페널티 회귀분석의 종류
    - 릿지 회귀분석: *ridge* regression analysis
    - 라소 회귀분석: *lasso* regression analysis
    - 일래스틱넷 회귀분석: *elasticnet* regression analysis

- 릿지 회귀분석: *ridge* regression analysis
  - 모델의 설명력에 기여하지 못하는 독립변수의 회귀계수 크기를
    - 0에 근접하도록 축소
  - *L2-norm* 페널티항으로 회귀모델에 페널티를 부과
    - L2-norm: 각 회귀계수의 제곱합
  - 릿지 회귀모델: 잔차의 제곱합과 L2-norm의 합을 최소화하는 회귀계수 추정
    - $\min_{\beta}\left[\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\lambda\sum_{j=1}^{p}\beta_j^{\ 2}\right]$
      - y_i: 관측값, $\hat{y}_i$: 예측값, n: 표본크기, p: 독립변수 개수, β_j: 회귀계수
      - λ: 페널티 튜닝 파라미터

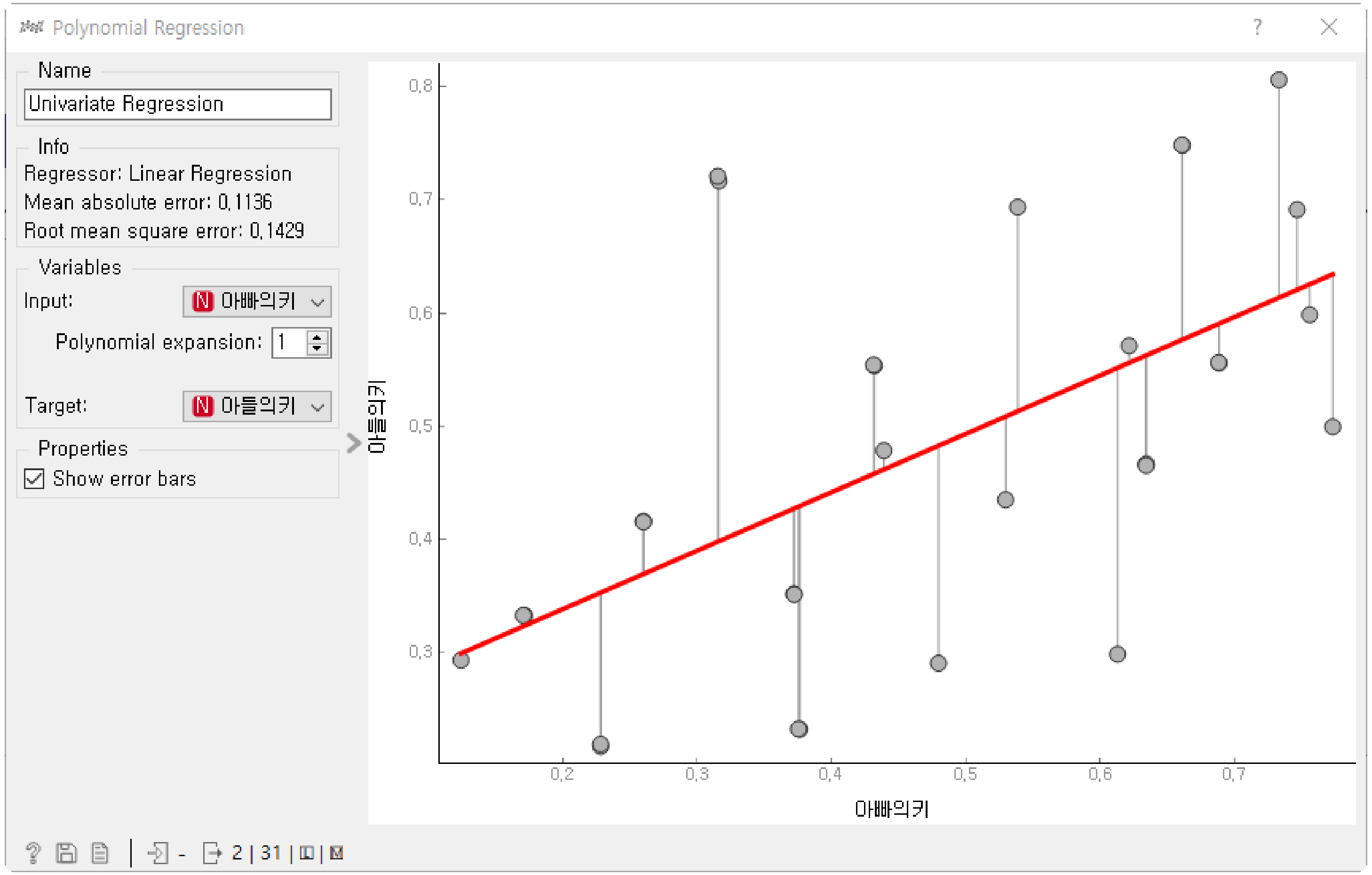
- **라소 회귀분석:** *lasso* regression analysis
  - 모델의 설명력에 기여하지 못하는 독립변수의 회귀계수 크기를
    - 0으로 만듬(해당 독립변수를 모델에서 제거)
  - *L1-norm* 페널티항으로 회귀모델에 페널티를 부과
    - L2-norm: 각 회귀계수의 절대값의 합
  - 라소 회귀모델: 잔차의 제곱합과 L1-norm의 합을 최소화하는 회귀계수 추정
    - $\min_{\beta}\left[\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right|\right]$
      - y_i: 관측값, $\hat{y}_i$: 예측값, n: 표본크기, p: 독립변수 개수, β_j: 회귀계수
      - λ: 페널티 튜닝 파라미터

- 엘라스틱넷 회귀분석: *elasticnet* regression analysis
  - *L1-norm*과 *L2-norm*을 모두 이용하여 회귀모델에 페널티를 부과
  - 엘라스틱넷 회귀모델:
    - $\min_{\beta}\left[\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\lambda\left\{(1-\alpha)\sum_{j=1}^{p}{\beta_j}^2+\alpha\sum_{j=1}^{p}\left|\beta_j\right|\right\}\right]$
      - y_i: 관측값, $\hat{y}_i$: 예측값, n: 표본크기, p: 독립변수 개수, β_j: 회귀계수
      - λ: 페널티 튜닝 파라미터, α: 모델의 혼합 정도를 통제하는 파라미터
    - $\alpha=0$: 순수한 릿지회귀모델
    - $\alpha=1$: 순수한 라소회귀모델
    - $0<\alpha<1$: 릿지회귀모델과 라소회귀모델의 혼합 정도를 통제

03. 선형 회귀

■ 회귀식의 기울기와 절편을 찾는 방법은?

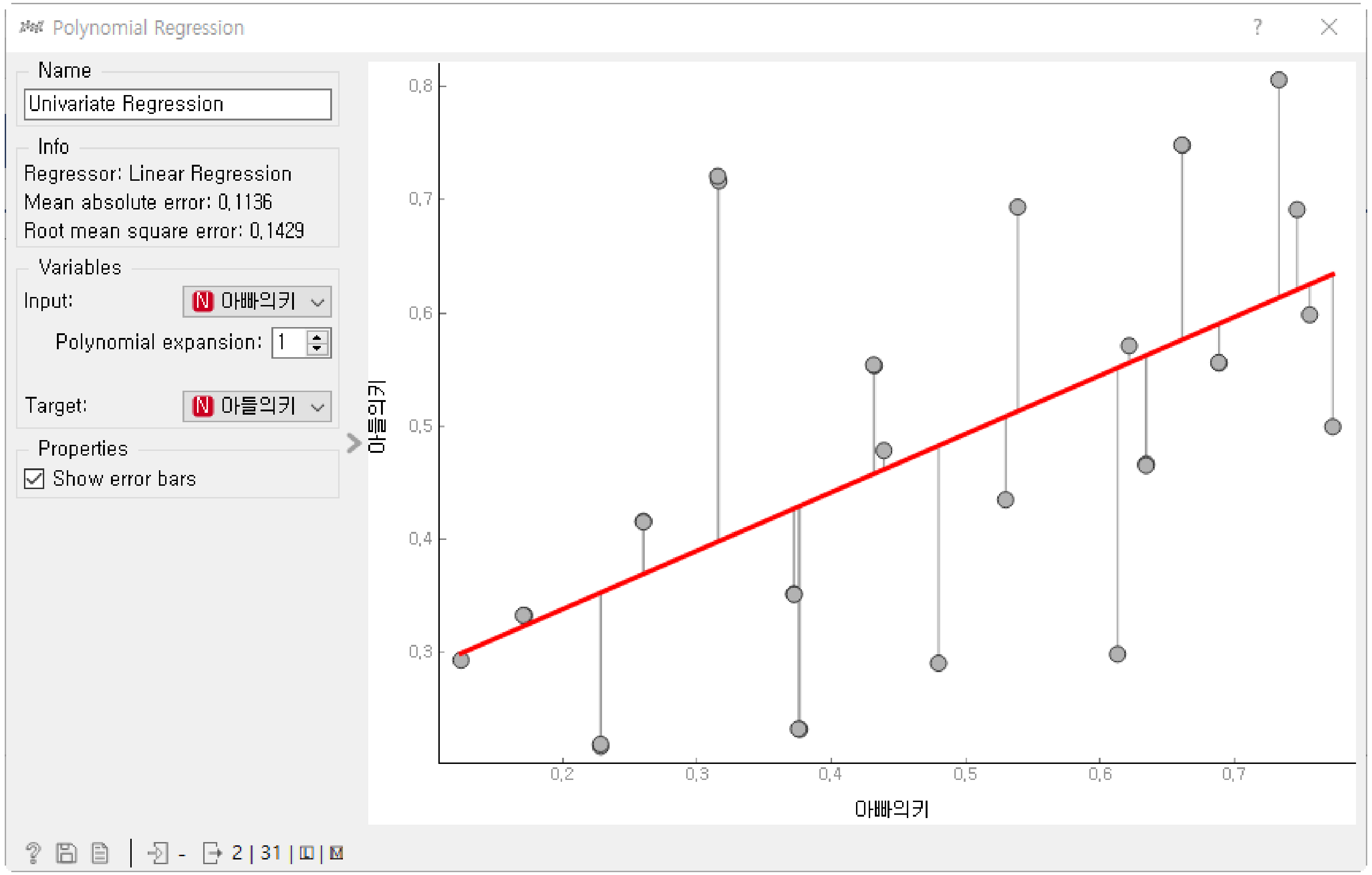
- 실제값과 예측값의 차이를 최소화하는 기울기와 절편 찾기



03. 선형 회귀

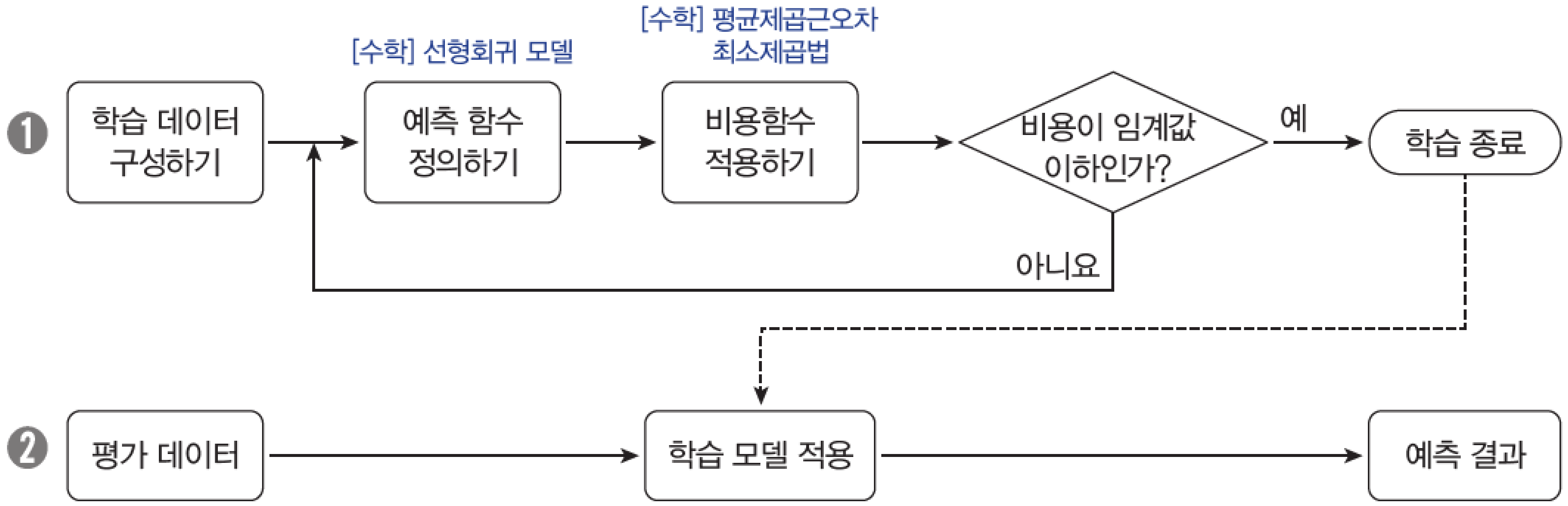
■ 회귀식의 기울기와 절편을 찾는 방법은?

- 실제값과 예측값의 차이를 최소화하는 기울기와 절편 찾기



- **비용 함수**: *Cost Function*
  - 실제값과 예측값의 차이를 측정하는 함수
    - 손실 함수: *Loss* Function
    - 목적 함수: *Objective* Function
  - 회귀식의 비용 함수를 최소화하는 기울기와 절편을 찾자.
    - 회귀식의 비용 함수: RSS, MAE, MSE, RMSE

▪ 선형 회귀식을 학습하기 위한 과정:



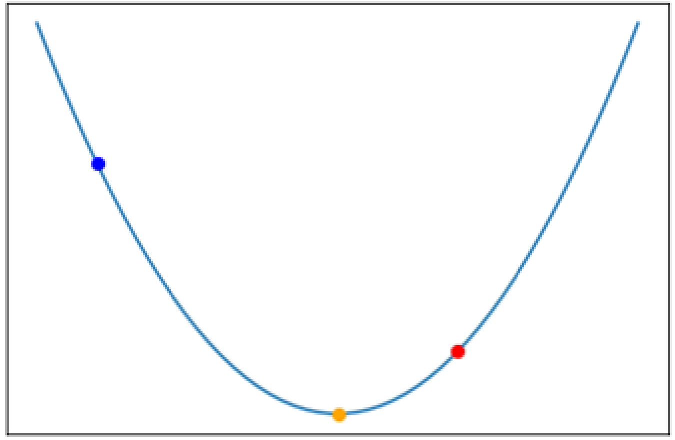
출처: 수학과 함께 하는 AI 기초, EBS

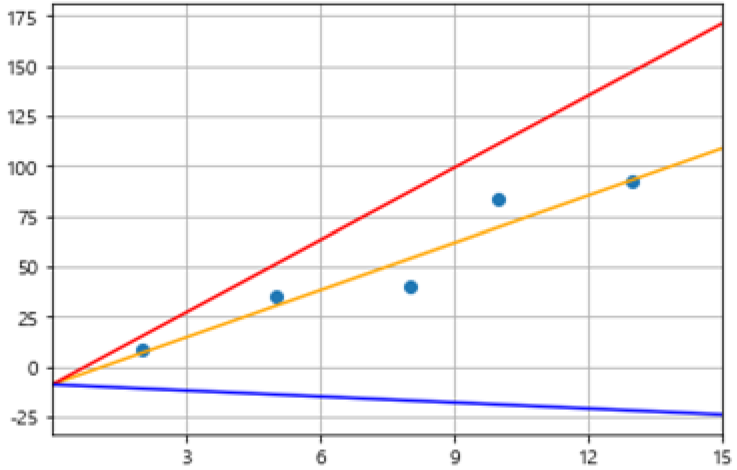
03. 선형 회귀

- **최소 제곱 추정법**: LSM, *Least Square Method*
  - 선형 회귀식의 기울기와 절편을 찾는 가장 일반적인 방법
  - 잔차: residuals
    - y의 실제값과 추정값 사이의 수직 거리의 차이: $r = y - \hat{y}$
  - 잔차 제곱합: RSS, *Residual Sum-of-Squares*
    - 잔차는 양수나 음수 모두 가능하므로 잔차의 제곱합을 구함
    - $RSS = \sum r^2 = \sum (y - \hat{y})^2$
  - 선형 회귀 분석의 목표:
    - 잔차 제곱합의 값이 최소가 되는 회귀식 찾기: $\hat{y} = \alpha\hat{x} + \beta$

경사하강법: *Gradient Descent*

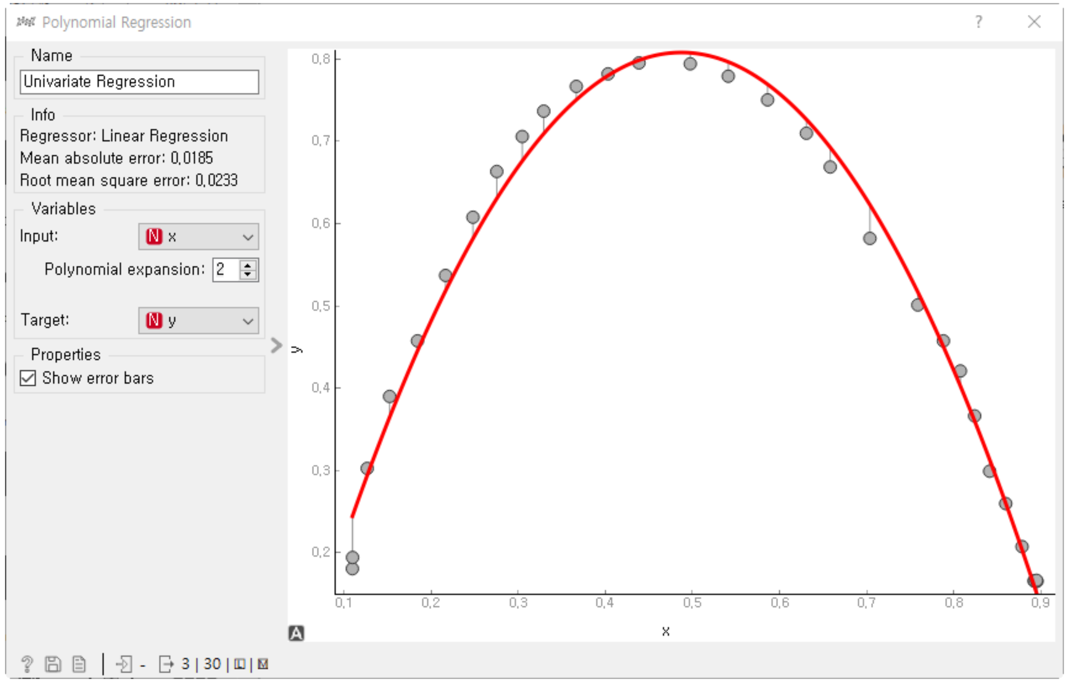
- 비용 함수(예: RSS)가 최소가 되는 기울기와 절편을 구하는 방법은?
  - 반복적인 계산을 통해 점진적으로 하강하면서 파라미터를 추정함
- 어떻게 오류가 작아지는 방향으로 파라미터를 보정할 수 있을까?

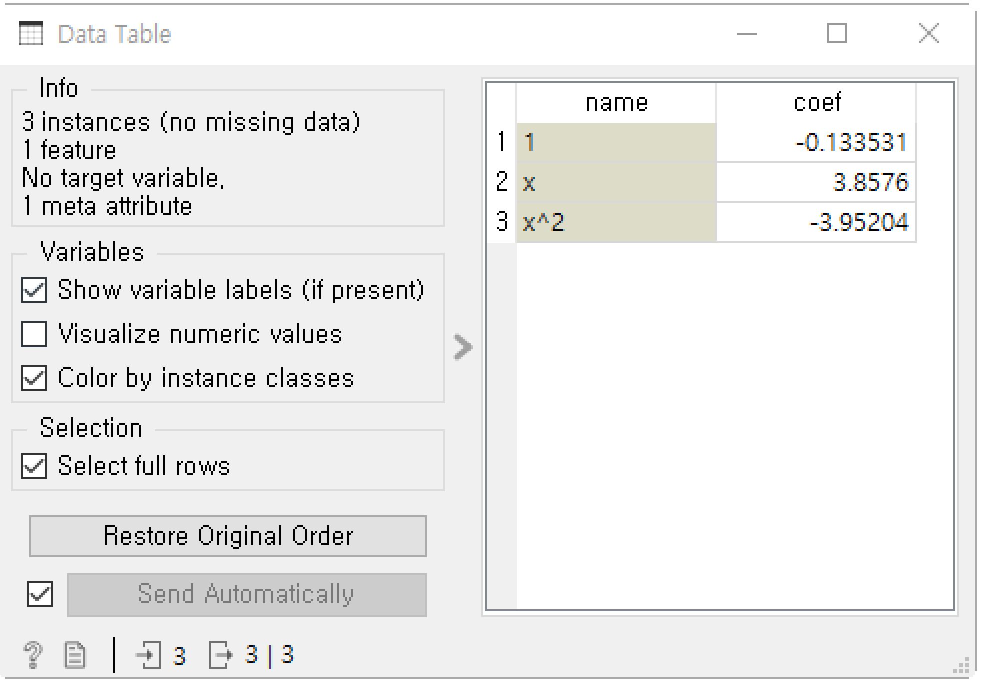




▪ 다항 회귀: *Polynomial* Regression

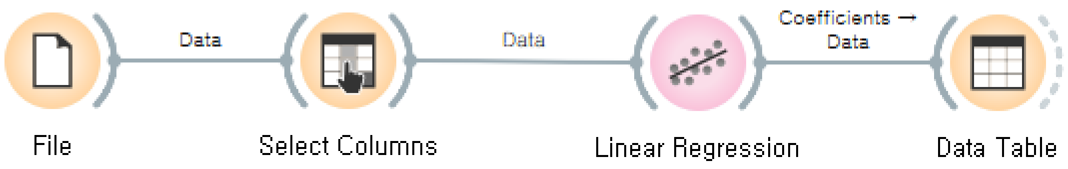
- 독립변수와 종속변수의 관계가 선형적일 때: $y = \alpha x + \beta$
- 만약, 두 변수의 관계가 2차 방정식, 3차 방정식의 관계라면?
  - $y = \beta_0 + \beta_1 x^1 + \beta_2 x^2 + \cdots + \beta_n x^n$

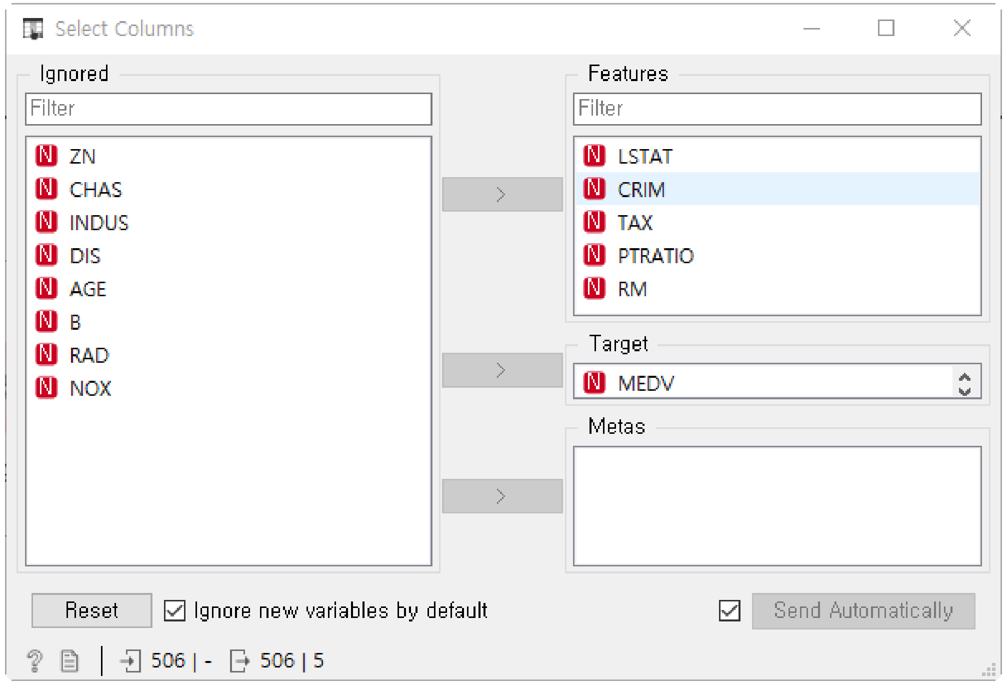


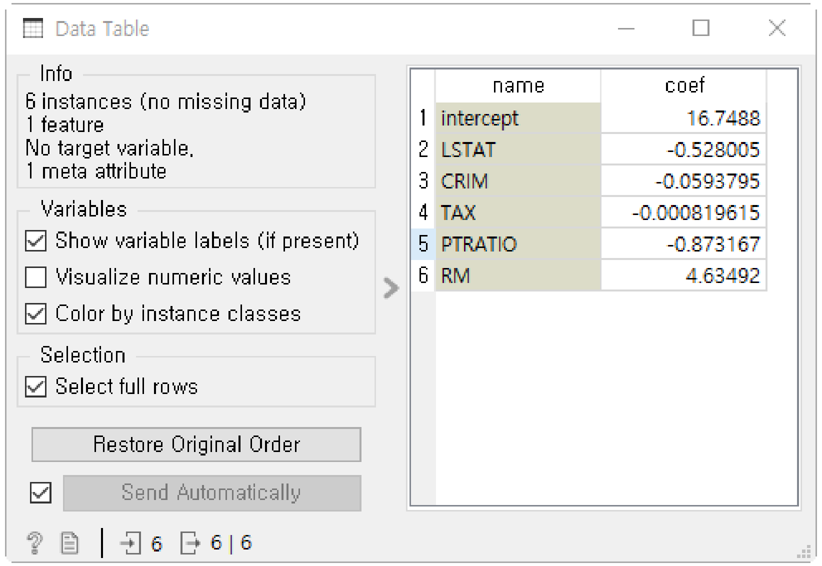


03. 선형 회귀

- **보스턴 주택 가격의 예측:**

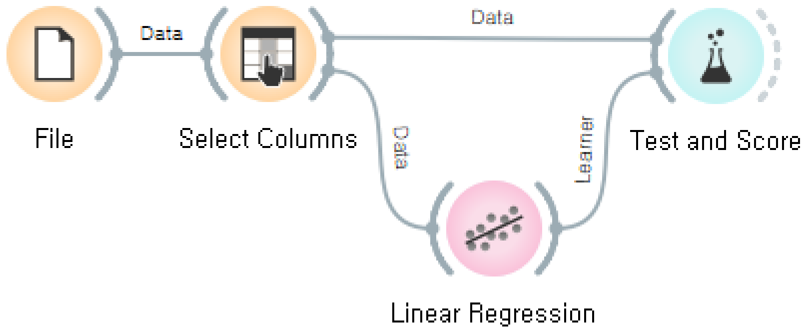


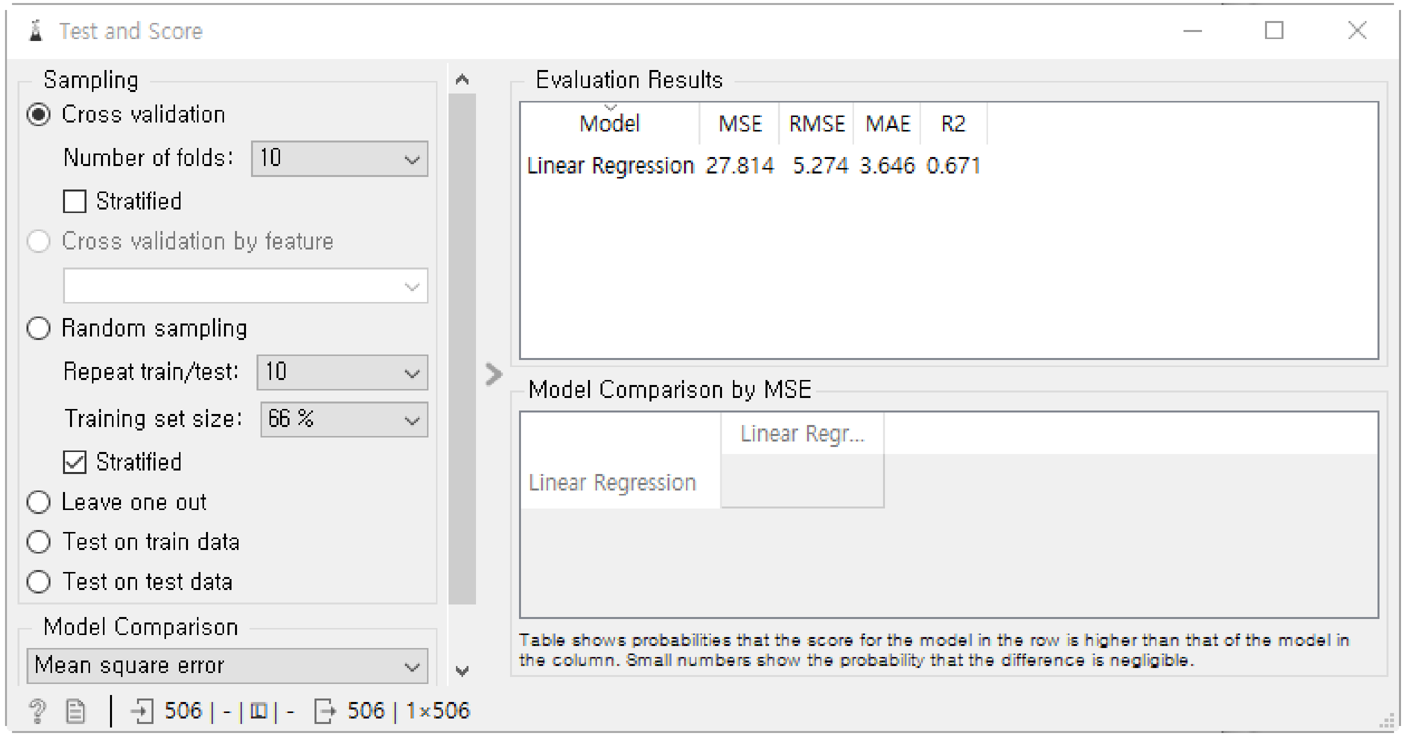




	name	coef
1	intercept	16.7488
2	LSTAT	-0.528005
3	CRIM	-0.0593795
4	TAX	-0.000819615
5	PTRATIO	-0.873167
6	RM	4.63492

▪ Orange: Test and Score





- 더 나은 성능을 가진 학습 모델을 만들려면?
  - 종속변수를 설명하는데 도움이 되는 독립변수가 여러 개일 때
    - 모든 독립변수가 종속변수를 설명하는데 동일하게 기여하는가?
    - 기여도가 높은 독립변수와 기여도가 낮은 독립변수를 구분
    - 기여도가 낮거나 거의 없는 변수들은 학습 모형에서 제외시킴
  - 설명변수의 숫자가 많을수록 좋은 학습 모델이라 할 수 있는가?
    - 설명변수의 숫자가 적을수록 좋은 학습 모델이라 할 수 있음

Any Questions?