회귀분석의 결정계수 (R-squared) 를 가장 쉽게 설명해 보자

x	y
10	30
20	40
30	50
40	80
50	90
60	100
70	120

위와 같은 x,y 데이터가 있습니다. 독립변수 x 에 따라 종속변수 y 가 변하는 이 데이터의 회귀모형, $y= \beta 0 + \beta 1x$ 을 구하고자 합니다. 그냥 통계 프로그램에 데이터를 넣으면 바로 값이 나오지만 원리를 아는 의미에서 수기로 먼저 계산해보도록 하겠습니다.

1) 상관계수 (correlation coefficient) 구하기

$$r = \frac{1}{n-1}\Sigma\left(\frac{x-\overline{x}}{s_x}\right)\left(\frac{y-\overline{y}}{s_y}\right)$$

x 의 평균: 40.0
x의 표준편차: 21.6
y 의 평균: 72.9
y 의 표준편차: 33.5

r= ((10-40.0) / 21.6 * (30-72.9) / 33.5 + (20-40.0) / 21.6 * (40-72.9) / 33.5 + (30-40.0) / 21.6 * (50-72.9) / 33.5 + (40-40.0) / 21.6 * (80-72.9) / 33.5 + (50-40.0) / 21.6 * (90-72.9) / 33.5 + (60-40.0) / 21.6 * (100-72.9) / 33.5 + (70-40.0) / 21.6 * (120-72.9) / 33.5) / (7-1) = 0.99

2) 회귀방정식 기울기 (β1) 구하기

```
β1 = r*Sy/Sx
```

β1 = 0.99 * (33.5/21.6) = 1.535

3) 회귀방정식 절편 (β0) 구하기

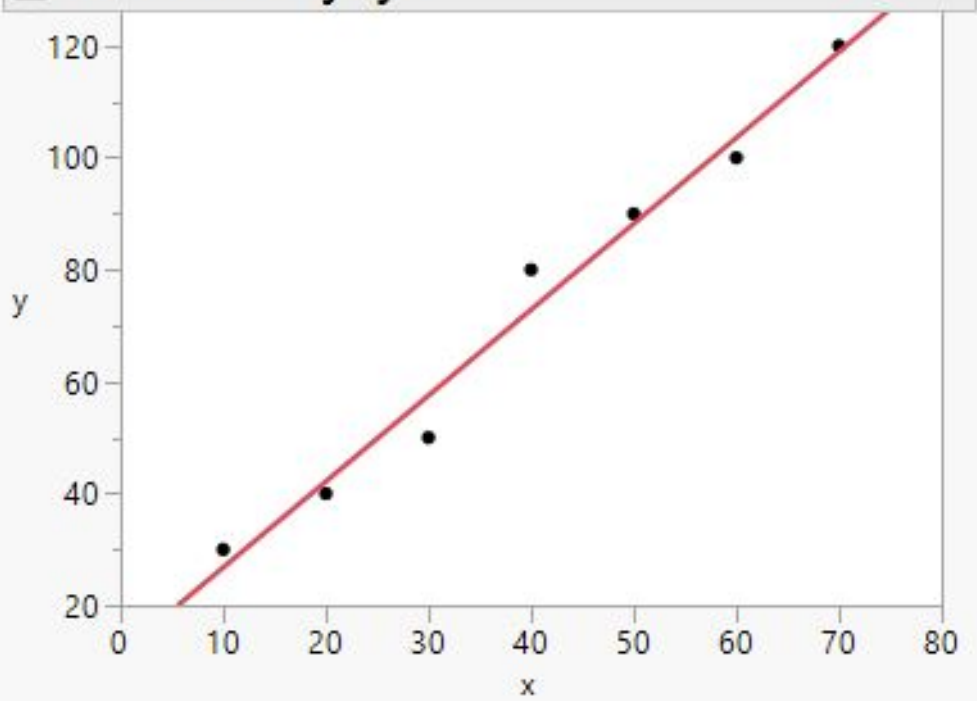
회귀모형, y= β0 + β1x 에서 ← 해당 회귀방정식은 특정 (x, y)를 지남.
72.9 = β0 + 1.535 * 40.0
∴ β0 = 11.5

자 그럼 수기로 계산된 회귀모형은 **y = 11.5 + 1.535x** 입니다.

이제 통계 프로그램으로 이 결과값이 맞는지 확인해 보겠습니다. 저는 JMP 를 사용합니다.

Untitled - Fit Y by X of y by x - JMP Pro

Bivariate Fit of y By x



Linear Fit

Linear Fit

y = 11.428571 + 1.5357143*x

Summary of Fit

RSquare	0.979343
RSquare Adj	0.975212
Root Mean Square Error	5.277987
Mean of Response	72.85714
Observations (or Sum Wgts)	7

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	6603.5714	6603.57	237.0513
Error	5	139.2857	27.86	Prob > F
C. Total	6	6742.8571		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	11.428571	4.460713	2.56	0.0505
x	1.5357143	0.099745	15.40	<.0001*

자!! 결과값이 y= 11.428571 + 1.5357143x 가 나왔습니다. 미세한 차이는 표준편차와 평균을 구할때 소수점 반올림에서 발생한 것입니다. 즉, 같은 결과값이라고 볼 수 있습니다.

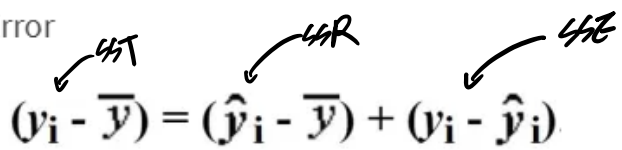
자!! 여기 통계 프로그램에서 결정계수 (R-squared) 를 이미 제공했습니다. 0.979343 이라고 합니다. 이렇게 결과를 얻었다고 여기서 멈추면 결정계수 (R-squared) 의 원리를 절대 이해 할수 없습니다. 이제 결정계수 (R-squared) 도 수기로 계산해 봅시다.

여기서 우리는 ANOVA 를 생각해야 합니다. 기본적으로 회귀분석을 하면 회귀방정식에 대한 통계적 유의성을 제공해 줍니다. 즉, 기울기가 0 이냐 아니냐를 기준으로 기울기가 0 이 아닐경우 그 회귀모형은 유의하다고 해석합니다.

하지만 ANOVA 는 각 데이터의 분산에 대한 해석입니다. ANOVA for Regression 의 개념을 지금부터 설명합니다.

```
SST = SSR + SSE
```

SST = Sum of Squares Total
SSR = Sum of Squares due to regression
SSE = Sum of Squared Error



어려워 보이는 개념 같지만 개념만 알면 초등학교 산수 문제 입니다. 우리의 x, y 값 옆에 회귀방정식에 의해서 예측된 값을 입력해 봅시다. JMP 결과값과 비교하기 위해서 JMP 에서 계산된 y=11.428571 + 1.5357143 x 를 사용합니다.

y=11.428571 + 1.5357143 x

x	yi	ŷi
10	30	26.8
20	40	42.1
30	50	57.5
40	80	72.9
50	90	88.2
60	100	103.6
70	120	118.9
	ȳ 72.86	

이제 이 값에서 여러가지 계산을 해 보겠습니다. 단순하게 **Data = Fit + Error** 라는 개념을 생각해 봅시다. 즉, 실제 개별 y 값에서 그 y 값 전체의 평균을 뺀값 (yi – ȳ) 을 개별 Data 라고 한다면, 그 값은 예측된 개별 y 값에서 실제 y 값 전체의 평균을 뺀 값 (ŷi -ȳ) + 실제 개별 y 값에서 예측된 개별 y 값을 뺀 값 (yi-ŷi) 과 같을 것입니다.

y=11.428571 + 1.5357143 x

$$(y_i - \overline{y}) = (\hat{y}_i - \overline{y}) + (y_i - \hat{y}_i)$$

x	yi	ŷi	Data =	Fit +	Error
10	30	26.8	-42.86	-46.07	3.21
20	40	42.1	-32.86	-30.71	-2.14
30	50	57.5	-22.86	-15.36	-7.50
40	80	72.9	7.14	0.00	7.14
50	90	88.2	17.14	15.36	1.79
60	100	103.6	27.14	30.71	-3.57
70	120	118.9	47.14	46.07	1.07
	ȳ 72.86				

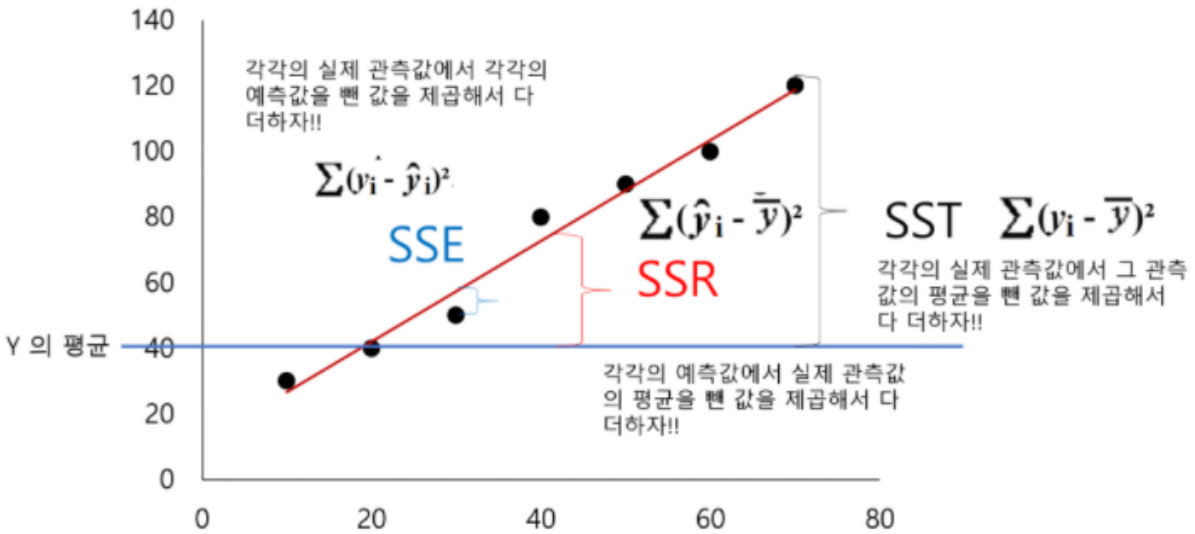
이렇게 계산된 값을 제곱해서 다 더해봅시다. 그러면 우리는 3개의 제곱합을 얻을수 있습니다. SST (6742.86) = SSR (6603.57) + SSE (139.29) 라는 값을 얻을수 있습니다.

y=11.428571 + 1.5357143 x

$$(y_i - \overline{y}) = (\hat{y}_i - \overline{y}) + (y_i - \hat{y}_i)$$

x	yi	ŷi	Data =	Fit +	Error	SST	SSR	SSE
10	30	26.8	-42.86	-46.07	3.21	1836.73	2122.58	10.33
20	40	42.1	-32.86	-30.71	-2.14	1079.59	943.37	4.59
30	50	57.5	-22.86	-15.36	-7.50	522.45	235.84	56.25
40	80	72.9	7.14	0.00	7.14	51.02	0.00	51.02
50	90	88.2	17.14	15.36	1.79	293.88	235.84	3.19
60	100	103.6	27.14	30.71	-3.57	736.73	943.37	12.76
70	120	118.9	47.14	46.07	1.07	2222.45	2122.58	1.15
	ȳ 72.86					6742.86	6603.57	139.29

이 계산식을 회귀 방정식 위에 표시하면 아래와 같습니다.



$$SST = SSR + SSE$$

이제 이 Sum of square 를 분산분석표로 옮겨 보겠습니다. 분산분석표에 들어갈 항목은 아래와 같습니다. 우리의 값을 심플하게 대입해 보겠습니다.

Source of variance

Source	DF	Sum of squared	Mean squared	F ratio
Model	1	SSR	SSR / 1	(SSR / 1) / (SSE / n - 2)
Error	n - 2	SSE	SSE / n - 2	
Total	n - 1	SST		

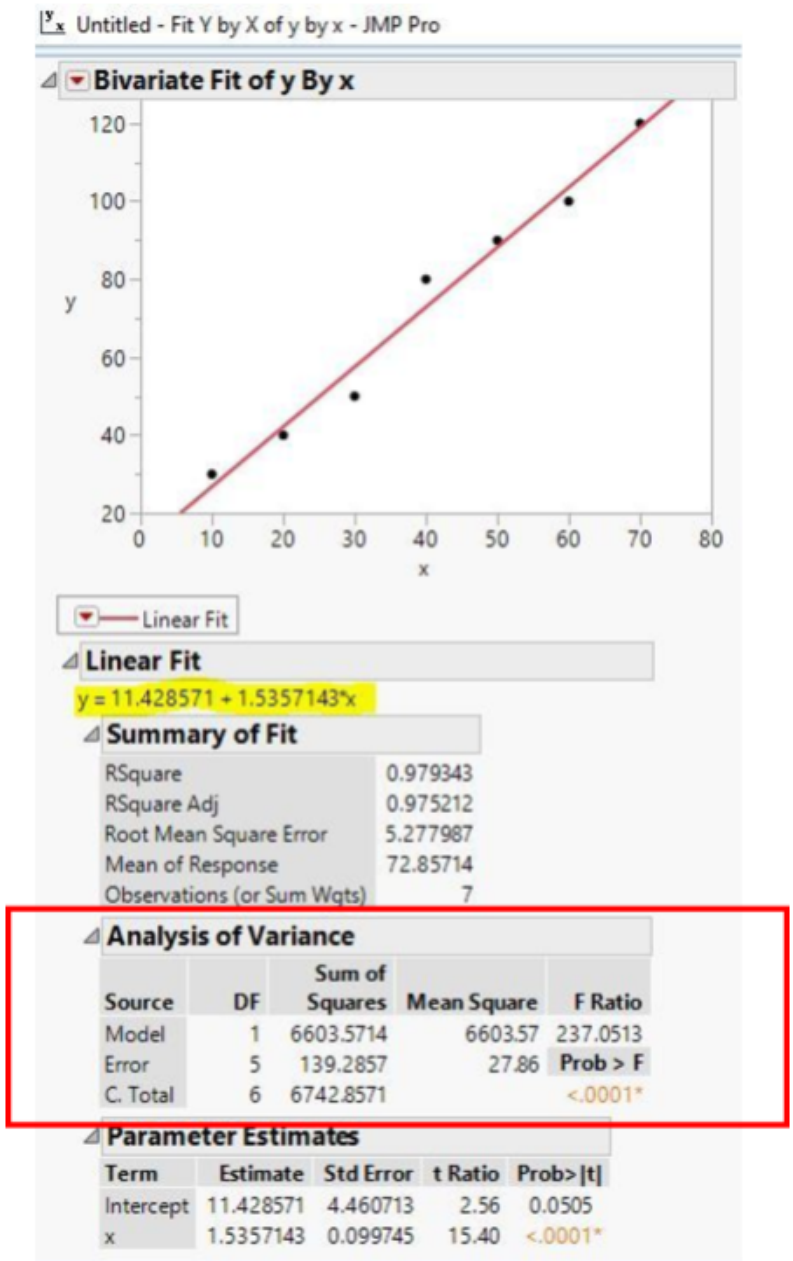
→ MSE

그러면 아래와 같은 분산분석표를 작성했습니다.

Source of variance

Source	DF	Sum of squared	Mean squared	F ratio
Model	1	6603.57	6603.57	237.051
Error	5	139.29	27.86	
Total	6	6742.86		

이제 우리가 수기로 작성한 분산분석표를 통계 프로그램의 결과와 비교해 보겠습니다.



같은 결과값 임을 알수 있습니다.

여기 까지 잘 따라 오셨다면 이제 결정계수 (R-squared) 를 계산하는 것은 식은죽 먹기 입니다.

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

그냥 SSR 에서 SST 를 나눠주면 됩니다.

R-squared = 6603.57 / 6742.86 = 0.979 로 계산됩니다. 위에 JMP 결과값을 보시면 같은 값임을 알수 있습니다. 즉, 결정계수는 전체 제곱합 중에서 회귀 제곱합이 차지하는 비율입니다. 이 비율이 높을수록 우리가 추정한 회귀 모형이 더 적합하다고 할수 있겠습니다.

↑ 'y' 편차의 합과 오편차의 합이 비슷할 수록 추정한 회귀 모형이 더욱 적합된 것임을 의미함.

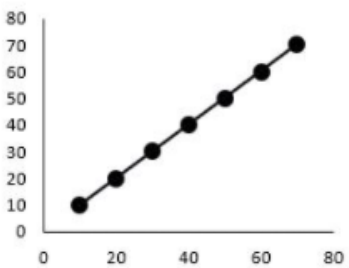
만일 y=x 의 모형일 경우 R-squared 는 어떻게 될까요? x 와 y 의 값이 같다면, 즉 y=x 라면 우리의 예측 y 값도 x 값과 같을 것입니다. 이 경우 Error 가 존재하지 않을 것입니다. 그리고 SST 와 SSR 은 같은 값일 것입니다. 그러면 R-squared 는 언제나 1 입니다. 즉, 회귀 방정식이 y=x 일 경우 R-squared 는 언제나 1 입니다.

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

x	yi	ŷi
10	10	10
20	20	20
30	30	30
40	40	40
50	50	50
60	60	60
70	70	70
	ȳ 40	

Data =	Fit +	Error
-30.00	-30.00	0.00
-20.00	-20.00	0.00
-10.00	-10.00	0.00
0.00	0.00	0.00
10.00	10.00	0.00
20.00	20.00	0.00
30.00	30.00	0.00

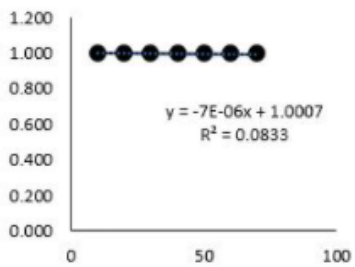
SST	SSR	SSE
900.00	900.00	0.00
400.00	400.00	0.00
100.00	100.00	0.00
0.00	0.00	0.00
100.00	100.00	0.00
400.00	400.00	0.00
900.00	900.00	0.00
2800.00	2800.00	0.00



반면 Y 값이 거의 같은 값일 경우는 어떨까요? x 에 반응하는 y 값이 거의 constant 한 경우, R-squared 는 극도로 낮아집니다. 만일 y 값이 완전 다 똑같다면 R-squared 는 계산되지 않습니다.

$$(y_i - \overline{y}) = (\hat{y}_i - \overline{y}) + (y_i - \hat{y}_i)$$

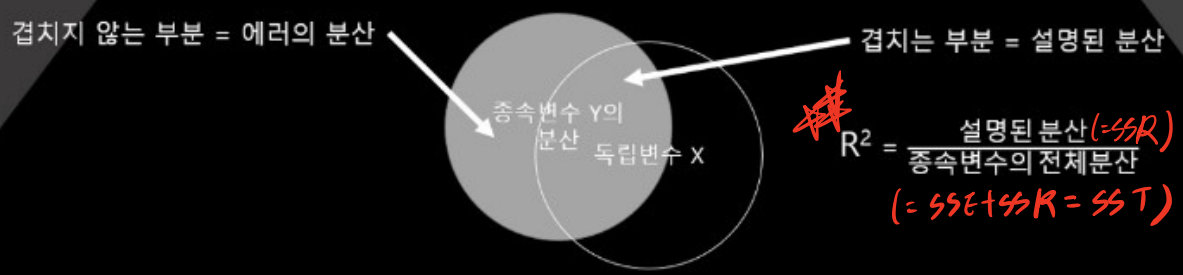
x	yi	ŷi	Data =	Fit +	Error	SST	SSR	SSE
10	1.000	1.00693	0.00	0.01	-0.01	0.0000002	0.0000423	0.0000480
20	1.001	1.00686	0.00	0.01	-0.01	0.0000003	0.0000414	0.0000343
30	1.001	1.00679	0.00	0.01	-0.01	0.0000003	0.0000405	0.0000335
40	1.000	1.00672	0.00	0.01	-0.01	0.0000002	0.0000396	0.0000452
50	1.001	1.00665	0.00	0.01	-0.01	0.0000003	0.0000387	0.0000319
60	1.000	1.00658	0.00	0.01	-0.01	0.0000002	0.0000378	0.0000433
70	1.000	1.00651	0.00	0.01	-0.01	0.0000002	0.0000370	0.0000424
	ȳ 1.000					0.0000017	0.0002772	0.0002786



R^2 란 무엇일까?

↑ "해당 회귀식이 종속변수를 얼마만큼 설명하는가?"에 대한 값

- 그러므로, 통계적 분석이란
  - 이론/논리를 통해서 종속변수를 설명할 수 있는 모델을 만들어
    - 이 모델에 들어가는 독립변수를 설정한 후
  - 종속변수의 분산을 모델(즉, 독립변수)로 설명하는데
  - 여기서 우리의 모델(즉, 독립변수)가 큰 문제가 없다면
    - 우리의 모델로 설명하고 남은 오차는 random한 오차임



$$R^2 = \frac{\text{설명된 분산}(=SSR)}{\text{종속변수의 전체분산}}$$

$(= SSE + SSR = SST)$

그렇다면

- R^2가 의미하는 것은 무엇이고 어떻게 해석해야 하나?
  - R^2는 모델의 분산 설명력이라고 볼 수 있음
  - 이는 우리가 만든 모델(즉, 독립변수)가 얼마나 데이터를 잘 설명했는지를 의미함
- R^2가 높으면 무조건 좋은 것인가?
  - 절대 그렇지 않음
  - 나름의 의미는 있으나, 높은 R^2가 모든 것을 완벽하게 하지는 못함
    - R^2 를 확인하기 전에 잔차도(residual plot)이 랜덤하게 분포함을 확인해야 함
    - 의미 없는 독립변수의 추가 조차도 R^2를 약간이라도 증가시킴
    - 그러나 독립변수의 추가는 자유도를 1 증가시켜 비용이 발생하는 것임
    - 높은 R^2는 과적합(overfitting)문제로 부터 자유롭지 않음