단순선형회귀모델

들어가기 전에...

분석 의뢰인이 아이스크림 가게를 운영하는 주인이라고 가정해보자.

판매용 아이스크림 주문 시, **예상되는 실제 판매량**만큼만 주문을 원한다.

이 때 만약 **평균 기온**을 활용하여 **미래 판매량**을 **예측**할 수 있다면?

들어가기 전에...

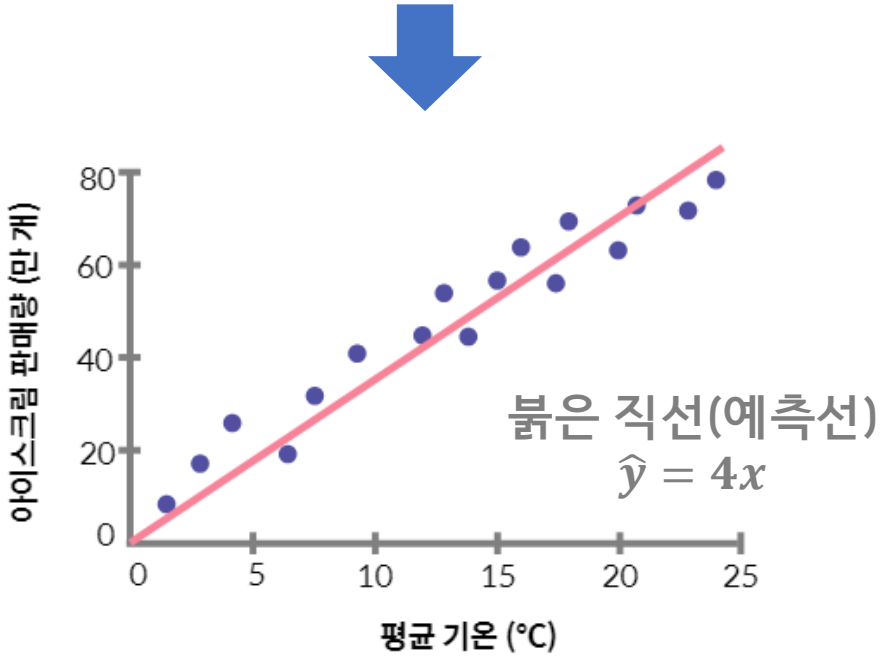
데이터: 6개월간의 일별 평균 기온과 일별 아이스크림 판매량

가정: 평균 기온과 판매량은 선형적인 관계를 가지고 있음

목표: 평균 기온에 따른 아이스크림 판매량 예측하기

X	Y
평균 기온(℃)	아이스크림 판매량(만개)
10	40
13	52.3
20	60.5
25	80

해결 방안: 단순선형회귀모델



들어가기 전에...

붉은 직선 $\hat{y} = 4x$ 의 의미를 잠깐 살펴보자

x: 평균 기온

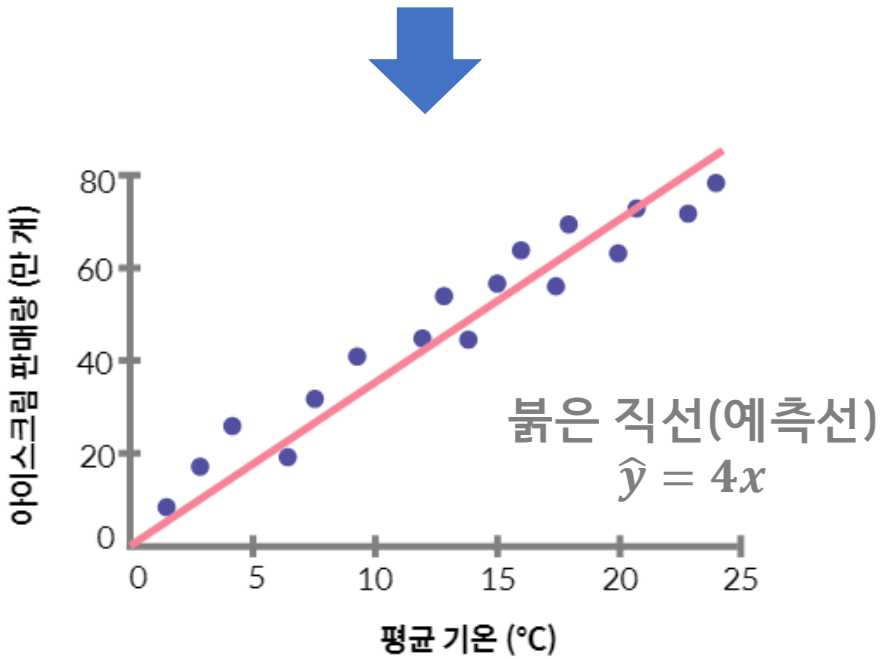
y:아이스크림 판매량

4: 기울기

그렇다면, $\hat{y}$은 무엇일까?

$\hat{y}$은 **선형회귀모델로 예측된 y값**이다.

X	Y
평균 기온(℃)	아이스크림 판매량(만개)
10	40
13	52.3
20	60.5
25	80



들어가기 전에...

$\underline{\hat{y} = 4x}$ 의미

평균 기온이 1만큼 커질때, 아이스크림 판매량의 예측값은 4만큼 증가함

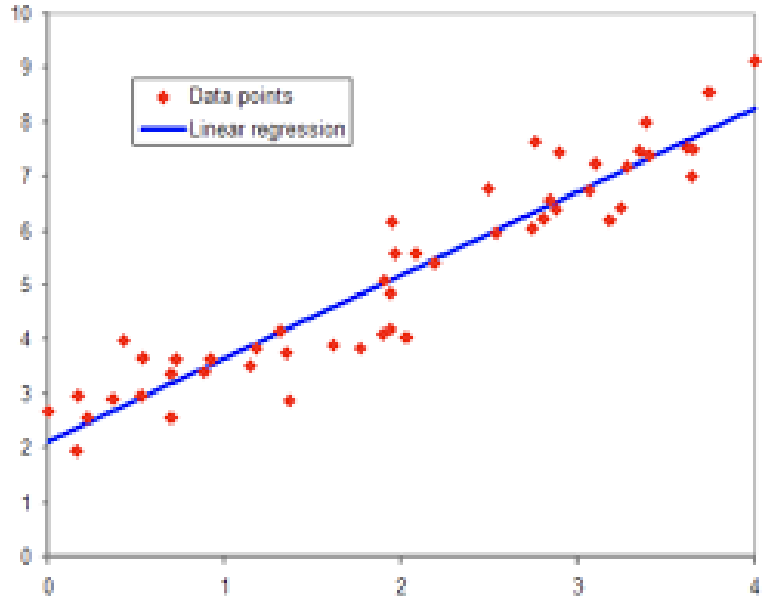
단순선형회귀모델(Simple Linear Regression)이란?

단순선형회귀모델은 하나의 X변수(독립변수)가 있는 선형회귀모델이다.

즉, Y변수(종속변수)를 하나의 독립변수로 최대한 정확하게 예측하는 선형 함수를 찾는다.

이것은 오른쪽 그림에 있는 직선과 같다. ($\boldsymbol{y = 2 + 3.5x}$라고 볼 수 있다.)

➡ 따라서 직선의 **최적 기울기**와 **절편**을 찾아야 된다!



단순선형회귀모델을 식으로 나타내면?!(1)

아래의 식은 **모집단 데이터**로 구축한 **단순선형회귀모델**이다.

$\boldsymbol{\alpha}$는 절편, $\boldsymbol{\beta}$는 기울기, $\boldsymbol{\varepsilon}$는 $\boldsymbol{y} - (\boldsymbol{\alpha} + \boldsymbol{\beta x})$이다. $(\boldsymbol{\alpha} + \boldsymbol{\beta x})$는 예측된 $\boldsymbol{y}$이다.

$$\boldsymbol{y = \alpha + \beta x + \varepsilon}$$

➡ **이것은 꽤나 이상적인 모델이다...**

우선, 샘플링한 데이터가 아니라 **모집단 데이터**(전국민의 건강 데이터)를 이용한다.

이후, 최적의 절편과 기울기를 계산한 선형회귀모델이다.

단순선형회귀모델을 식으로 나타내면?!(2)

아래의 식은 **모집단 데이터**로 구축한 **단순선형회귀모델**이다.

$\boldsymbol{\alpha}$는 절편, $\boldsymbol{\beta}$는 기울기, $\boldsymbol{\varepsilon}$는 $\boldsymbol{y-(\alpha+\beta x)}$이다. $(\boldsymbol{\alpha+\beta x})$는 예측된 $\boldsymbol{y}$이다.

$$\boldsymbol{y=\alpha+\beta x+\varepsilon}$$

➡ **ε(오차)는 무엇일까?**

우리는 선형회귀모델로 실제 y값을 100% 예측할 수 없다.

그래서 실제 y값과 모집단 데이터로 예측한 y값$(\alpha+\beta x)$에 차이가 있고, 그것이 오차이다.

단순선형회귀모델을 식으로 나타내면?!(3)

아래의 식은 **샘플링한 데이터**를 활용한 **단순선형회귀모델**이다.

$\hat{\alpha}$는 추정된 절편, $\hat{\beta}$는 추정된 기울기, ϵ는 $y-(\hat{\alpha}+\hat{\beta}x)$이다. $(\hat{\alpha}+\hat{\beta}x)$는 예측된 y이다.

$$y=\hat{\alpha}+\hat{\beta}x+\epsilon$$

➡ **왜 "추정된"이라는 단어를 쓸까?**

실제 절편과 기울기는 모집단 데이터를 활용해서 구한다.

우리는 모집단에 대해 조사할 수 없기 때문에 샘플을 이용해 절편과 기울기를 추정하는 것이다.

단순선형회귀모델을 식으로 나타내면?!(4)

아래의 식은 **샘플링한 데이터**를 활용한 **단순선형회귀모델**이다.

$\hat{\alpha}$는 추정된 절편, $\hat{\beta}$는 추정된 기울기, ϵ는 $y-(\hat{\alpha}+\hat{\beta}x)$이다. $(\hat{\alpha}+\hat{\beta}x)$는 예측된 y이다.

$$y=\hat{\alpha}+\hat{\beta}x+\epsilon$$

➡ **ϵ(잔차)는 무엇일까?**

우리는 선형회귀모델로 실제 y값을 100% 예측할 수 없다.

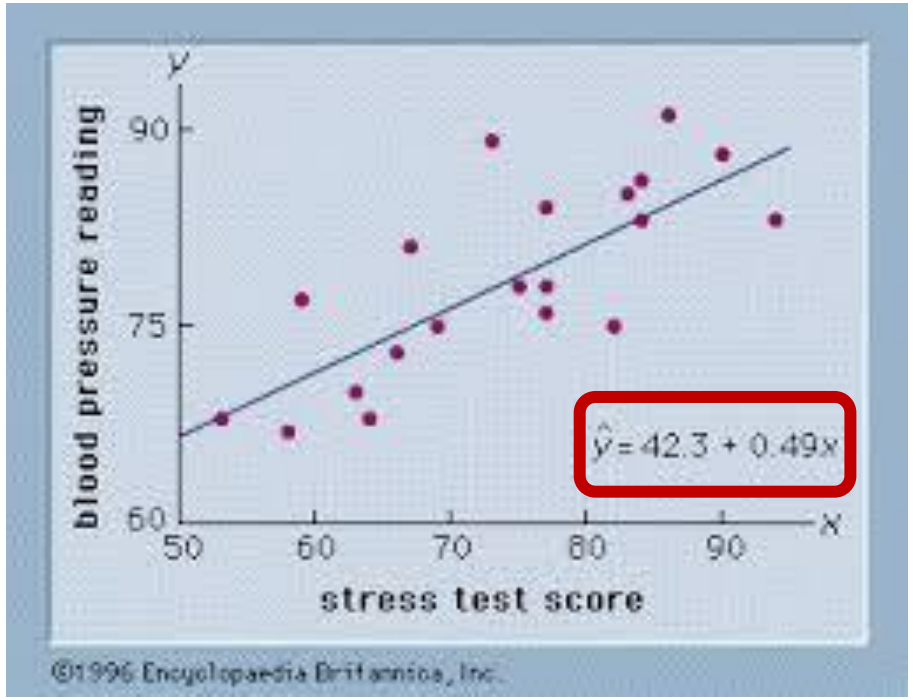
그래서 실제 y값과 샘플 데이터로 예측한 y값$(\hat{\alpha}+\hat{\beta}x)$에 차이가 있고, 그것이 잔차이다.

선형회귀모델을 직선그래프로 나타내자

아래의 식은 **샘플링한 데이터**를 활용한 **단순선형회귀모델**이다.

$\hat{\alpha}$는 추정된 절편, $\hat{\beta}$는 추정된 기울기, $\hat{y}$은 예측된 y값이다.

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$



선형회귀모델 해석

들어가기 전에...

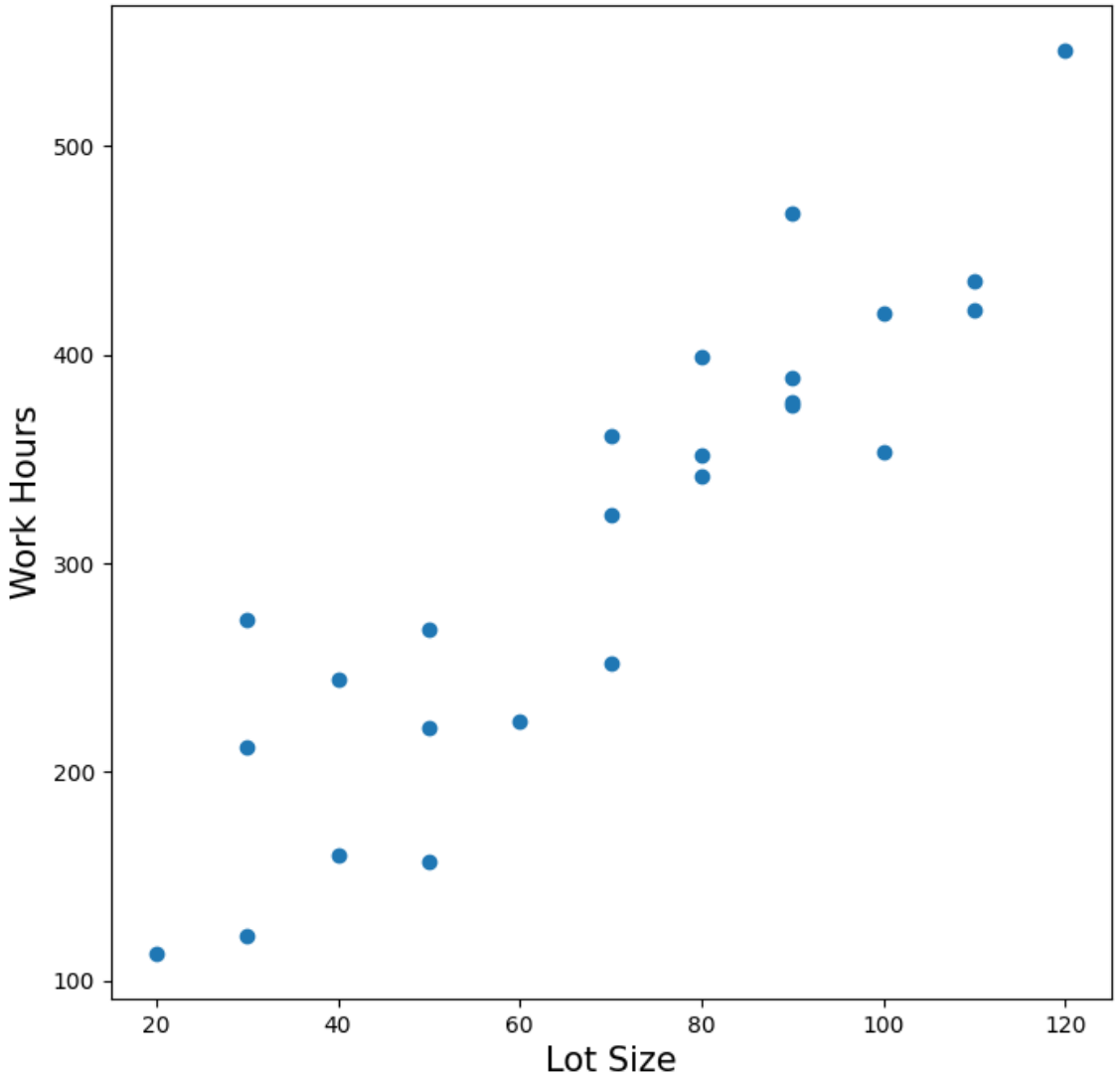
제품의 크기에 따른 작업 시간을 예측하고자 한다.

우리는 단순선형회귀모델을 이용할 것이다.

변수는 아래와 같이 정의한다.

1) Lot Size: 제품의 크기(X변수)

2) Work Hours: 작업 시간(Y변수)



선형회귀모델은 간단한 코드 몇줄로 구축이 가능하다.

구축했다치고, 결과를 해석해보자!!!

선형회귀모델 결과표

Dep. Variable:	Work_hours	**R-squared:**	0.822
Model:	OLS	**Adj. R-squared:**	0.814
Method:	Least Squares	**F-statistic:**	105.9
Date:	Mon, 24 Jul 2023	**Prob (F-statistic):**	4.45e-10
Time:	18:31:42	**Log-Likelihood:**	-131.64
No. Observations:	25	**AIC:**	267.3
Df Residuals:	23	**BIC:**	269.7
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	62.3659	26.177	2.382	0.026	8.214	116.518
Lot_size	3.5702	0.347	10.290	0.000	2.852	4.288

Omnibus:	0.608	**Durbin-Watson:**	1.432
Prob(Omnibus):	0.738	**Jarque-Bera (JB):**	0.684
Skew:	0.298	**Prob(JB):**	0.710
Kurtosis:	2.450	**Cond. No.**	202.

괭장히 복잡하다..

하지만 우리는 R-squared와

Coef와 P>|t|만 알면 된다!

R-Squared란 무엇인가?

선형회귀모델에서 X변수가 Y변수를 얼마나 잘 설명해주는지 보여주는 지표이다.

0과 1사이에 있으며, 1에 가까울수록 모델의 설명력이 좋다고 판단한다.

R-Squared란 무엇인가?

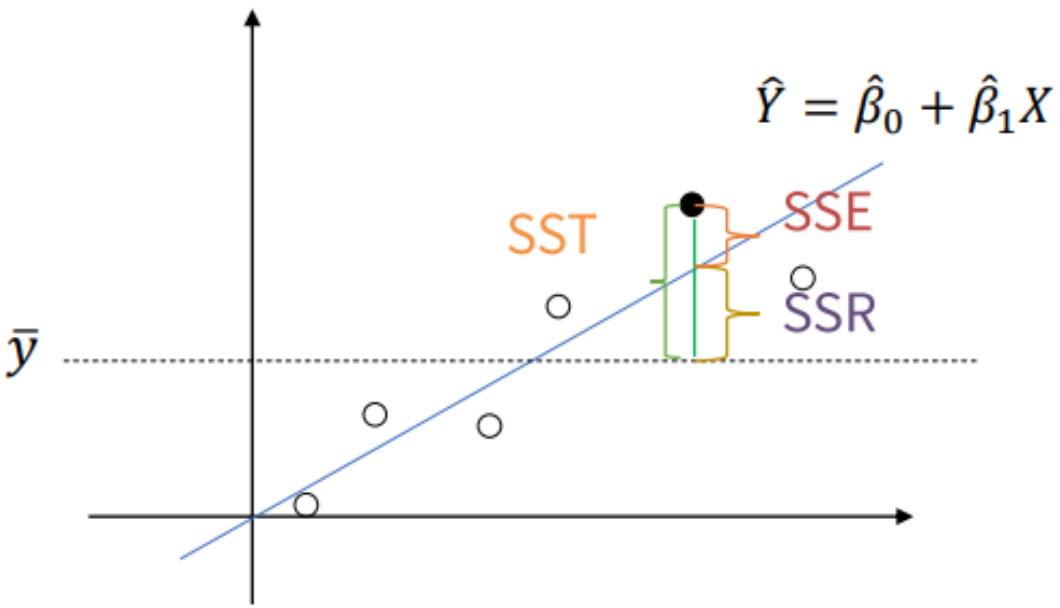
이번에는 $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$ 직선이 있다.

$\bar{y}$는 Y의 평균이다.

사실 평균은 누구나 구할 수 있다.

평균만으로 Y를 예측하는 것은 **불가능**하다.

정확한 예측을 위해 $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$ 을 만들었다.



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

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{SSE}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{SSR}}$$

SST (Total sum of squares) SSE (Error sum of squares) SSR (Regression sum of squares)

R-Squared란 무엇인가? - SST

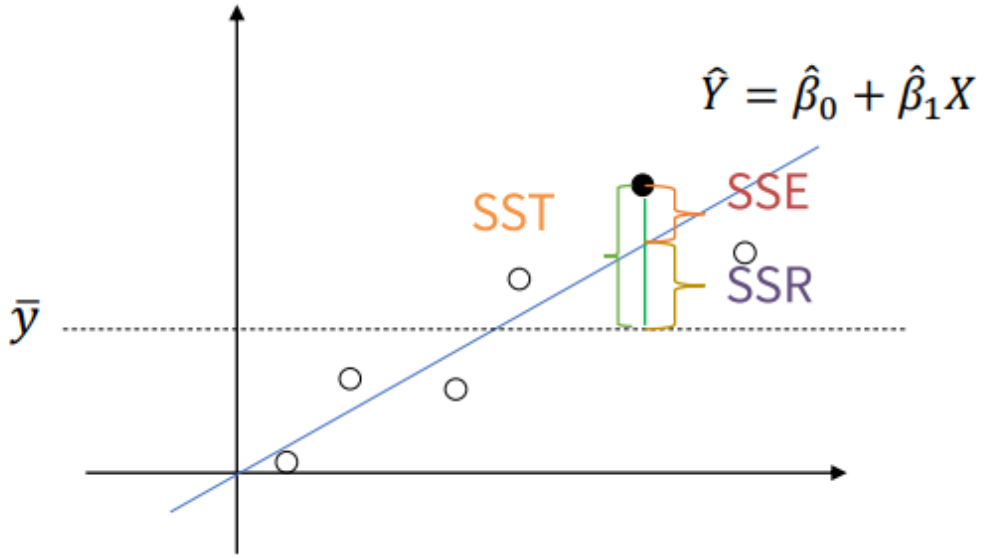
R-Squared를 이해하려면 **SST, SSE, SSR**을 알아야 된다.

SST

y의 평균과 실제 y의 차이에 대한 수식

꽤 차이가 날 것이다.

평균만으로 전체 y를 예측하는 것은 불가능하다.



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

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{SST (Total sum of squares)}} = \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{SSE (Error sum of squares)}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{SSR (Regression sum of squares)}}$$

R-Squared란 무엇인가? - SSE

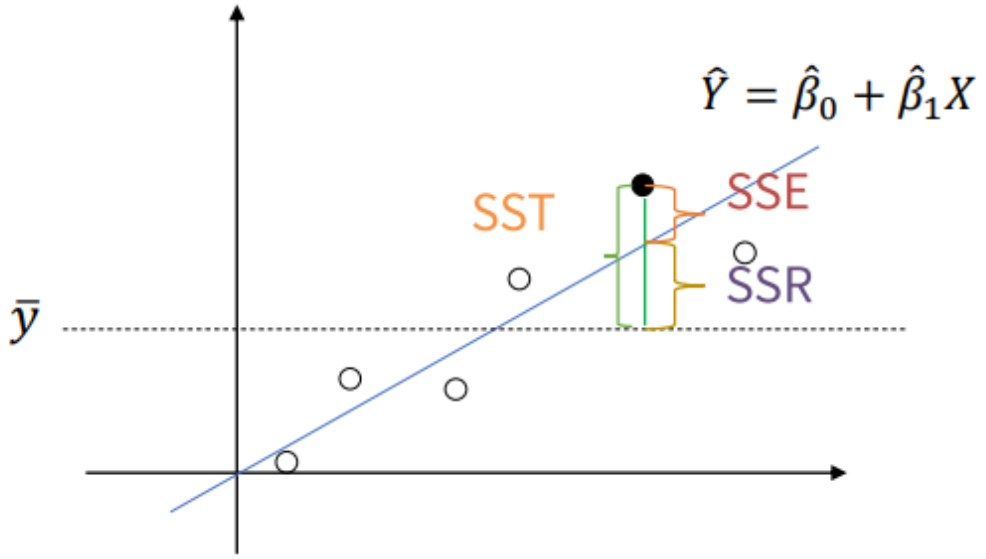
R-Squared를 이해하려면 **SST, SSE, SSR**을 알아야 된다.

SSE

실제 y와 예측된 y의 차이에 대한 수식

선형회귀모델로는 완벽하게 y를 예측할 수 없다.

설명할 수 없는 변동이라고도 부른다.



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

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{SST (Total sum of squares)}} = \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{SSE (Error sum of squares)}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{SSR (Regression sum of squares)}}$$

R-Squared란 무엇인가? – SSR

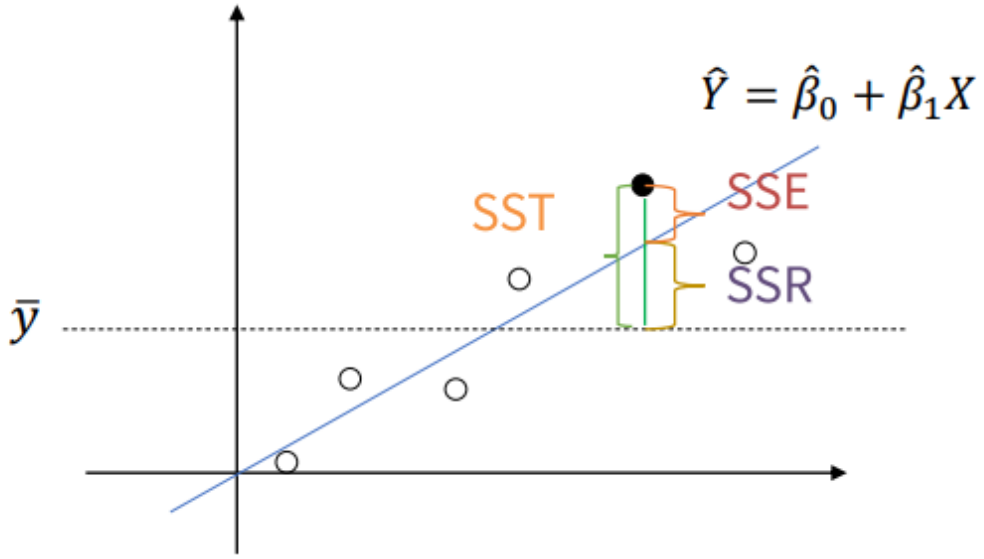
R-Squared를 이해하려면 **SST, SSE, SSR**을 알아야 된다.

SSR

예측된 y와 평균 y의 차이에 대한 수식

SST에서 SSE를 뺀 나머지이자,

설명할 수 있는 변동이라고도 부른다.



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

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{SST (Total sum of squares)}} = \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{SSE (Error sum of squares)}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{SSR (Regression sum of squares)}}$$

R-Squared란 무엇인가? – 수식

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

Coefficient란 무엇인가?

Coefficient는 선형회귀모델에서 **X 변수가 1 증가시 Y 변수의 평균이 얼마나 변하는지**를 나타낸다.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	62.3659	26.177	2.382	0.026	8.214	116.518
Lot_size	3.5702	0.347	10.290	0.000	2.852	4.288

예를 들면, 앞에서의 예시에서 Lot_Size의 회귀계수는 3.57이다.

Lot_Size가 1 증가했을 때 Work_hours의 평균이 **3.5702 증가함**을 의미한다.

잠깐!!! Y변수의 평균과 $\hat{y}$은 같다!!!

P-Value란 무엇인가?

P-Value(유의확률)은 **샘플 데이터로 구한 모델이 모집단에도 적용 가능한지 가늠하는 지표**이다.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	62.3659	26.177	2.382	0.026	8.214	116.518
Lot_size	3.5702	0.347	10.290	0.000	2.852	4.288

➡ **특정 X변수의 P-Value가 0.05 보다 작은 경우 어떻게 해석할까?**

모집단에 대한 가설[특정 X변수는 Y변수에 영향을 주지 않는다]을 거부할 증거를 제공한다.

따라서 특정 X변수는 통계적으로 유의하며 선형회귀모델에 추가할 가치가 있다.