Finance Analytics
Chapter3. Linear Regression Model
Part 4. Goodness of Fit and Prediction

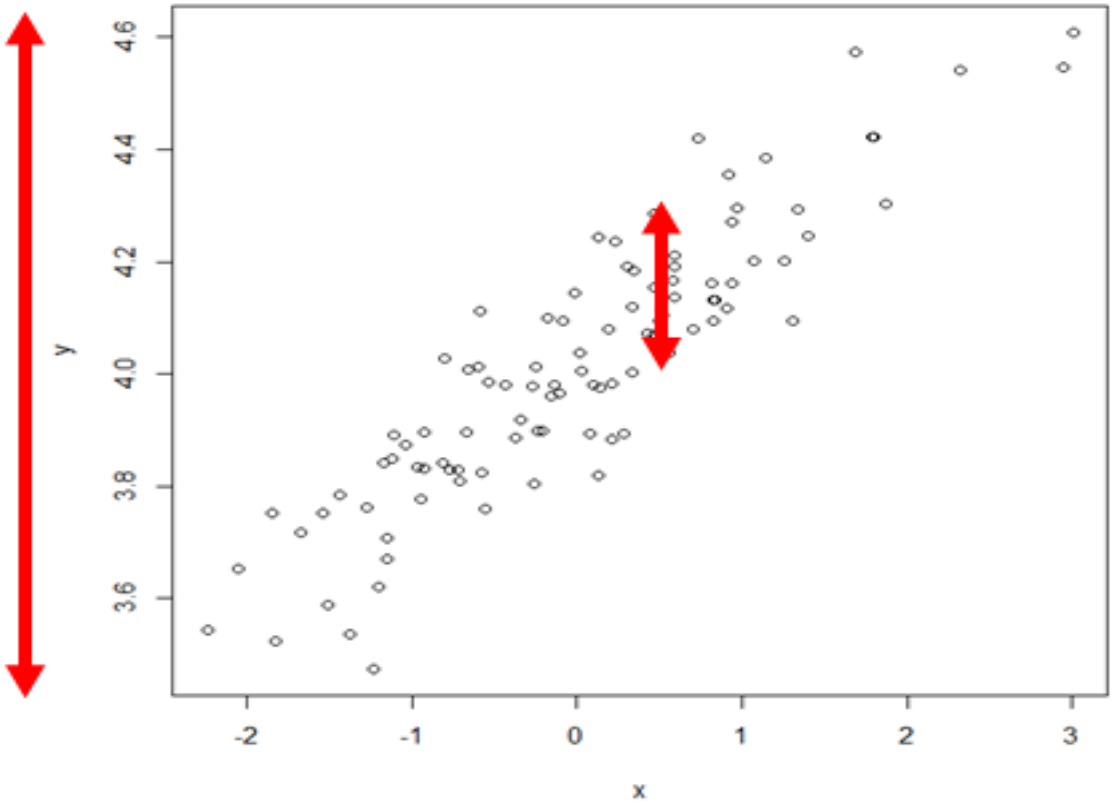
권태연

한국외대 국제금융학과

선형회귀모형의 적합도(설명력) 판단

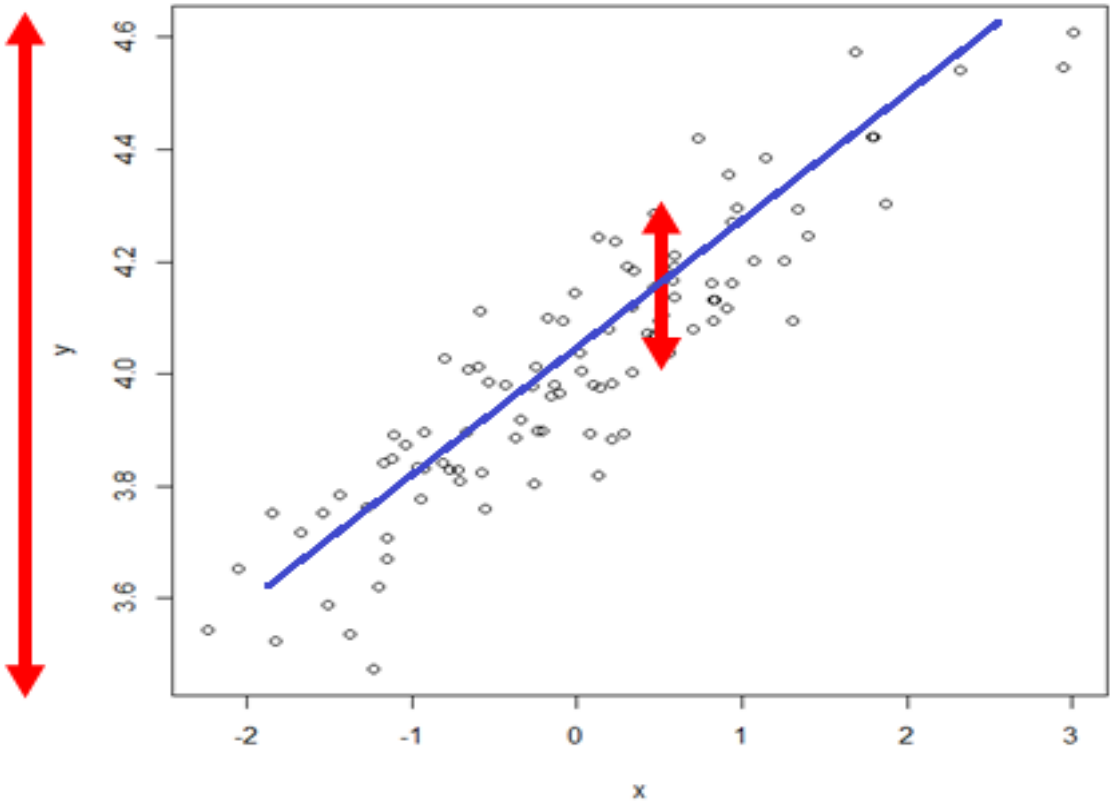
R-square, 결정계수

- R^2로 표기, 결정계수 (coefficient of determination)
- 회귀선의 적합도 (goodness of fit)의 측정
- 종속변수(Y)의 변이 중 독립변수(X)들이 설명하는 비율



R-square, 결정계수

- R^2로 표기, 결정계수 (coefficient of determination)
- 회귀선의 적합도 (goodness of fit)의 측정
- 종속변수(Y)의 변이 중 독립변수(X)들이 설명하는 비율



R-square, 결정계수

- Y의 총변동(SST, Sum of Squared Total or TSS(total sum of square)) $\sum(y_i - \bar{y})^2$은 다음과 같은 두 부분으로 나뉜다.
  1. 설명된 제곱합
     독립변수를 이용하여 설명할 수 있는 변동, 체계적부분의 변동
  2. 설명되지 않는 제곱합 (비체계적 부분의 변동)
     $= \sum(y_i - \hat{y})^2$

1. 설명된 제곱합

   : $\sum(\hat{y} - \bar{y})^2$
   : SSR (Sum of Squared Regression)
   : ESS (Explained Sum of Square)

2. 설명되지 않는 제곱합

   : SSE (Sum of Squared Error)
   : RSS (Residual Sum of Square)

R-square, 결정계수

$$\begin{aligned} \sum(y_i - \bar{y})^2 &= \sum(\hat{y} - \bar{y})^2 + \sum(y_i - \hat{y})^2 \\ SST &= SSR + SSE \end{aligned}$$

이고 이때 결정계수 R^2는 다음과 같다.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- 단순선형회귀모형에서 결정계수 R^2은 상관계수 r^2값과 같다.
- 결정계수의 해석 : 독립변수 X들은 종속변수 Y를 $R^2 \times 100$ % 설명한다.
- 시간당 임금함수 예제의 결정계수 확인 및 해석
- 높은 R^2가 적합도가 높은 모형이지만, X변수의 갯수가 늘어남에 따라 R^2는 증가, 이에 Adjusted R-square 이용

R-square, 결정계수

- 높은 R^2가 적합도가 높은 모형이지만, X변수의 갯수가 늘어남에 따라 R^2는 증가, 이에 Adjusted R-square 이용
- 독립변수의 갯수만큼 페널티(penalty)부여

F-test

$\hat{y} = b_0 + b_1 X_1 + b_2 X_2 + ... + b_k X_k$

- t-검정: $H_0 : B_k = 0$에 대한 테스트: 단일 회귀계수에 대한 유의성 검정
- F-검정: $H_0 : B_1 = B_2 = ... = B_k = 0$에 대한 테스트: 모든 회귀계수에 대한 유의성 검정
- F-검정에서 귀무가설 Reject의 의미?
- F-검정에서 귀무사설 Not reject의 의미?

$$F = \frac{MeanSSR}{MeanSSE} = \frac{SSR/df1}{SSE/df2}$$

- F-검정통계량 값이 크면 기각. (유의수준에 따라서 결정됨)
- 시간당 임금함수 예제의 F-test결과 확인 및 해석

Prediction and Model validation
모형의 예측력과 적정성

Goodness of fit of Model vs Validation of Model

- Prediction ($\hat{y}$):
적합도가 높은 모형의 추정 후, 독립변수(X 들)값이 주어져 있을때 Y 값의 예측
$\hat{y} = b_0 + b_1 X_1 + b_2 X_2 + ... + b_k X_k$

- How well can we explain or predict the dependent variable based on the independent variables ?

- Goodness of Fit (설명력) : assess how well model fit the data.
- Validation, Prediction Performance, Predictive power (예측력)
: 모형의 적용가능성, 해석에 차이

data= training set+ validation set

- Training Set : 모형 적합(model fitting, model estimation)을 위해 사용되는 자료의 부분
- Model Validation Set: 적합된 모형으로 "prediction performance"를 보기 위해 사용되는 자료의 부분
- over-fitting, 모형의 과적합 문제.