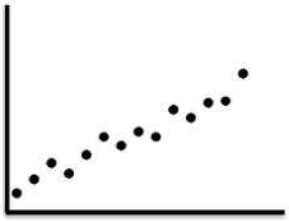
7-3. 결정계수 그리고 설명된 편차와 설명되지 않은 편차

① 결정계수 R^2은 회귀식이 얼마나 정확한지를 나타내는 숫자이다.

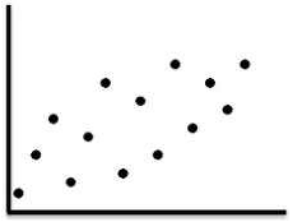
② 회귀분석에서 회귀식을 활용하여 무언가를 예측할 때, 정답인 실제값이 아닌 틀릴 확률이 존재하는 예측값이 나오면서 항상 오차가 발생한다.

㉠ 오차가 작다 = 점들이 모여 있는 밀도가 높다 = 회귀식의 정확도가 높다

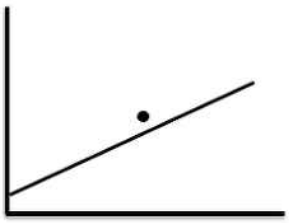
㉡ 오차가 크다 = 점들이 모여 있는 밀도가 낮다 = 회귀식의 정확도가 낮다



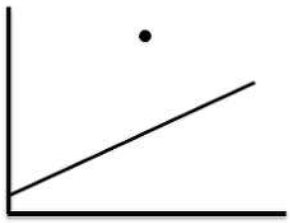
밀도가 높다



밀도가 낮다

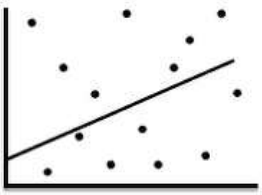


오차가 작다
<회귀식의 정확도가 높다>

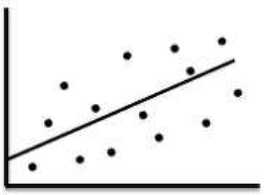


오차가 크다
<회귀식의 정확도가 낮다>

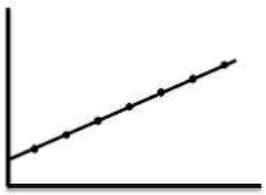
③ R^2값이 1에 가까운 큰 값일수록 완벽한 회귀모형이다.



$R^2=0$
<믿을 게 못 된다>



$R^2=0.5$
<어느 정도 믿을 만 한다>



$R^2=1$
<믿을 만 하다>

④ R^2 결정계수(상관비) = $\frac{\text{설명된 변화량}}{\text{총변화량}}=\frac{SSR}{SST}$

㉠ Y의 총제곱합(SST) 중 추정된 회귀식이 설명하는 변동량(SSR)의 비율이다.

㉡ 적합모형에 의해 설명된 변동의 비율에 대한 측도이다.

㉢ 총변화량 중 추정된 회귀식이 설명하는 변동량(SSR)의 비중이 크면 클수록 회귀식이 원래의 자료를 잘 정리 요약하여 반영한다.

㉣ 상관계수의 제곱 r^2이다.

㉤ 일반적으로는 100을 곱하여 %단위로 쓴다.

㉥ 사회과학 연구에서는 R^2값이 70%만 넘어도 상당히 큰 값으로 생각한다.

〖예〗 중학교 5명의 영어점수의 총편차제곱합은 10, 설명된 편차제곱합은 9.0 그리고 설명되지 않은 편차제곱합은 1.0이다.

㉠ 국어, 영어점수에서의 결정계수 = 설명된 변화량/총변화량 = 9/10=0.9

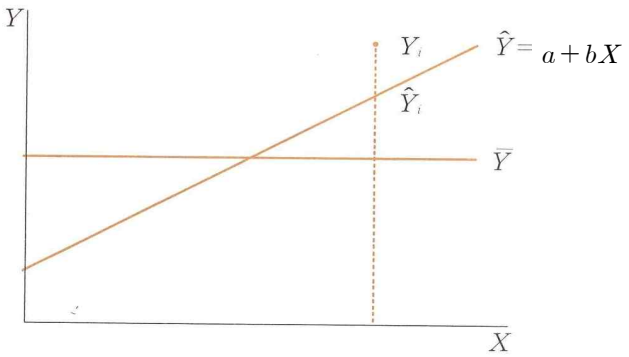
㉡ Y변수의 총변화량의 90%를 X변수가 설명해 줌.

㉢ X변수의 총변화량의 90%를 Y변수가 설명해 줌.
(상관관계는 상호관계로 해석하기 때문)

㉣ 전체 변동의 90% 가량이 회귀직선에 의해 설명되므로 회귀직선은 유용하다고 판단된다.

⑤ 결정계수 구하기 : '설명된 편차' 와 '설명되지 않은 편차'

설명된 편차	설명되지 않은 편차
같은 X_i값을 가진 모든 대상은 각기 다른 Y_i값을 가졌다 하여도 기대되는 $\hat{Y}_i$값은 같으므로 X_i값이 같을 때 $(\hat{Y}_i - \overline{Y})$값은 항상 같다.	X_i값이 같으므로 그에 대응하여 기대되는 $\hat{Y}_i$값은 같으나 각 개인의 Y_i값이 다르므로 $(Y_i - \hat{Y}_i)$값은 다르다. 이는 개인차 혹은 측정의 오차 등에 의하여 발생된다.



㉠ $Y_i = \overline{Y} + (\hat{Y}_i - \overline{Y}) + (Y_i - \hat{Y}_i)$
= 'Y변수를 대표하는 평균' + '예견되는 기댓값과 평균의 차이'
+ '관찰된 점수와 기댓값의 차이'

㉡ 각 개인의 총편차 $Y_i - \overline{Y}$ = $(\hat{Y}_i - \overline{Y})$ + $(Y_i - \hat{Y}_i)$
설명된 편차 (explained deviation) 설명되지 않은 편차 (unexplained deviation)

㉢ 편차합 = 0, 따라서 편차의 제곱의 합을 이용

㉣ 총변화량은 설명된 변화량과 설명되지 않은 변화량으로 구분된다.

$$SST = \sum(Y_i - \overline{Y})^2 \quad SSR = \sum(\hat{Y}_i - \overline{Y})^2 \quad SSE = \sum(Y_i - \hat{Y}_i)^2$$

총편차제곱합 (total sum of squares) + 설명된 편차제곱합 (regression sum of squares) = 설명되지 않은 편차제곱합 (residual sum of squares)

총제곱합

7-4. 회귀의 선형성에 관한 검정

① X에 관한 Y의 모회귀직선의 방정식 $Y=\beta_0+\beta_1 X$에서 $\beta_1=0$이면, X의 어떤 값에 대해서도 $Y=\beta_0$로 일정하므로 X에 대한 Y의 값을 추정할 수 없으며 이 방정식은 회귀성을 갖지 않는다고 한다. 따라서 회귀성의 유무에 대한 검정, 즉 $\beta_1=0$에 대한 검정을 해야 한다.

② 표본회귀직선의 방정식 $\hat{Y}=b_0+b_1 X$와 모회귀직선의 방정식 $Y=\beta_0+\beta_1 X$에서 $H_0:\beta_1=0,\ H_1:\beta_1\neq 0$을 검정하는데 분산분석을 이용한다.

③ 회귀직선의 유의성 검정은 다음의 F통계량을 이용한다.

$$F=\frac{S_R^2}{S_E^2}\sim F(1,\ n-2)$$

㉠ 회귀평균제곱(regression mean square) $S_R^2=\dfrac{SSR}{1}$

㉡ 잔차평균제곱(residual mean square) $S_E^2=\dfrac{SSE}{n-2}$

㉢ 유의수준 α에서 $F>F_\alpha(1,\ n-2)$이면 귀무가설 H_0를 기각한다.

④ 회귀직선의 유의성 검정을 위한 분산분석표는 다음과 같다.

요인	제곱합	자유도	평균제곱	F	R^2
설명된 변화량(회귀선)	SSR	1	$S_R^2=SSR$	$F=\dfrac{S_R^2}{S_E^2}$	$\dfrac{SSR}{SST}$
설명되지 않은 변화량(잔여분)	SSE	$n-2$	$S_E^2=\dfrac{SSE}{(n-2)}$		
총변화량	SST	$n-1$			

〖예〗 다음 표는 기억력 X와 판단력 Y를 조사하여 얻은 자료이다. X에 관한 Y의 회귀직선의 방정식을 구하고 회귀직선의 유의성을 검정하시오.

X	8	9	9	10	10	10	11	12	12	15
Y	2	3	4	3	5	7	6	7	8	9

[풀이]

㉠ 가설 세우기 $H_0:\beta_1=0,\ H_1:\beta_1\neq 0$

㉡ 분산분석표

요인	제곱합	자유도	평균제곱	F	임계값
회귀	38.55	1	38.55	38.55/1.23 = 31.34	$F_{0.01}(1,\ 8)=11.26$
잔차	9.85	8	9.85/8=1.23		
총합	48.40	9			

㉢ $F=31.34>11.26$이므로 $\alpha=0.01$에서 귀무가설을 기각한다.