1. 회귀분석 개념

(1) 종속변수(목표변수 output, Y)와 설명변수(예측변수 input, X, 개수는 p) 간 함수관계(통계적 모형)를 도출하여 $Y = f(X_1, X_2, \ldots, X_p) + e$ 예측변수 관측값에 의해 목표변수 값을 추정한다.

(2) 오차항 e가 없다면 함수는 수확적 결정함수이다. (수학함는 예측변수 값이 주어지면 목표변수 값이 단일 값으로 결정된다)

(3) 통계적 모형 : $Y = f(X_1, X_2, \ldots, X_p) + e$ 에서 관심 분석 내용

- 고려된 모든 설명변수$(X_1, X_2, \ldots, X_p)$ 중 종속변수를 유의하게 설명하는(영향을 주는) 변수??? 유의성 검정
- 유의한 설명변수 중 종속변수에 영향을 가장 많이 미치는 (중요도) 변수 도출
- 설명변수의 값이 주어졌을 때 종속변수의 예측값(fitted value)과 신뢰구간

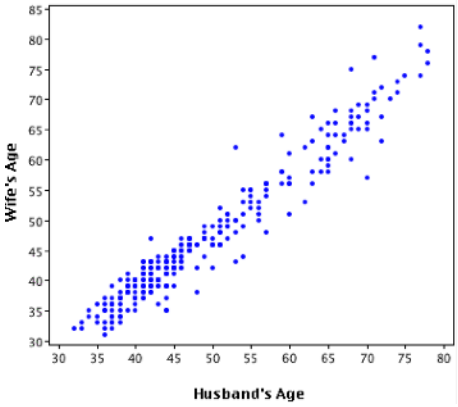
함수 형태 중 선형을 가장 선호 $Y_i = a + b_1X_{1i} + b_1X_{2i} + \ldots + b_1X_{pi} + e_i$

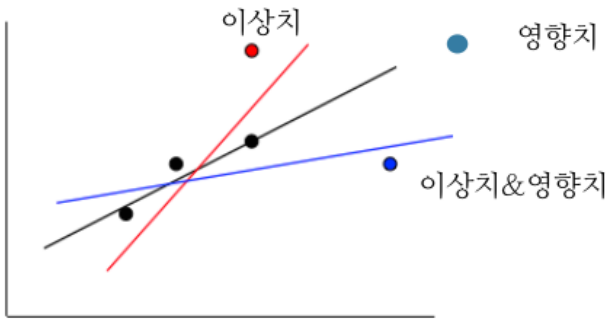
- 해석이 용이 : 모수(회귀계수 $b_1, b_2, \ldots$)는 기울기, 설명변수 한단위 증가(감소)할 때 종속변수의 단위 변화량
- 로그변환을 통하여 선형으로 변환 가능 : Cobb-Douglas 생산함수 $Q = \alpha K^{\beta} L^{\lambda} u$ (Q=생산량, K=자본, L=투입노동, α, β, λ 모수), 인구성장모형 $P_t = \alpha e^{\beta T} e_t$ (P=인구수, T=시간, α, β 모수) 모두 양변에 로그를 취하여 선형함수로 만들 수 있음
- 오차항은 $e_i \sim N(0, \sigma^2)$을 가정한다.
- 회귀분석을 분산분석모형과 함께 선형모형으로 불리는 것은 예측변수의 함수 형태가 선형(직선)이기 때문이다.

2. 산점도 SCATTER PLOT

(1) 개념

- 두 변수의 함수관계를 이차원 그래프로 표현, y-축 : 종속변수, x-축 : 설명변수





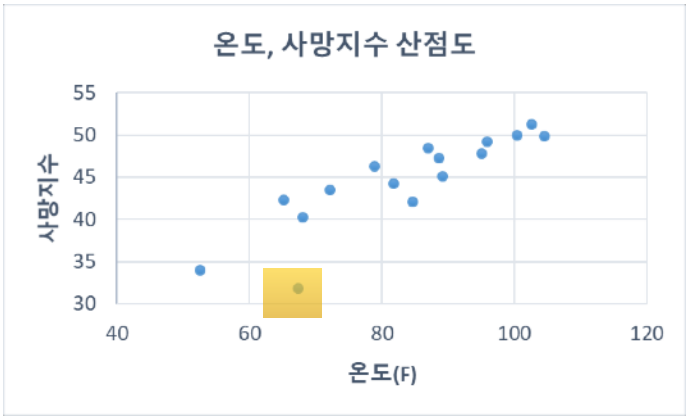
(부부의 나이 함수 관계)

(2) 활용

- 함수 관계 진단 : 선형이 아닌 경우 회귀분석 적용을 위하여 선형 변환이 필요
- 함수관계에서 이상치, 영향치 진단 - 잔차분석에서 문제 해결

사례 데이터 : 💾 CANCER.csv

연 평균 온도(F: Fahrenheit, 설명변수)가 여성 종양 사망지수(mortality index)에 영향을 미치는지 알아보기 위하여 유럽 17개 도시를 대상으로 조사한 자료이다.



온도가 올라갈수록 암 사망지수는 직선적으로 증가하는 경향을 보인다.

계열 "mortality" 요소 "67.3" (67.3, 31.8) (노랑 사각형) 관측치는 이상 관측치로 판단됨. 다른 관측치에 비해 사망지수가 낮은 지역임

3. 단순회귀 모형 SIMPLE REGRESSION MODEL

(1) 단순 회귀모형

종속변수 Y , 설명변수 X라 하고 첨자(subscript) i는 관측치를 표현하며 데이터 개수를 n 이라 하면

$$y_i = \alpha + \beta x_i + e_i, \; i = 1,2,...,n$$

- α : 회귀계수(regression coefficient), 모수, 절편(intercept)
- β : 회귀계수, 설명변수 X의 기울기, 설명변수 X가 한 단위 증가할 때마다 종속변수 Y의 증가량(미분 계수), 다중회귀모형에서는 편미분 계수
- x : 설명변수이며 결정변수 deterministic (즉 확률변수가 아님)
- e_i : 오차항(error term), 회귀직선($y_i = \alpha + \beta x_i$)에 의해 설명되지 못하는 부분, 오차항이 없으면 통계모형이 아니라 수학모형

(2) 가정 assumption

(a) 모수인 회귀계수 α, β 는 unknown but fix 미지이지만 고정 - 확률변수 아님

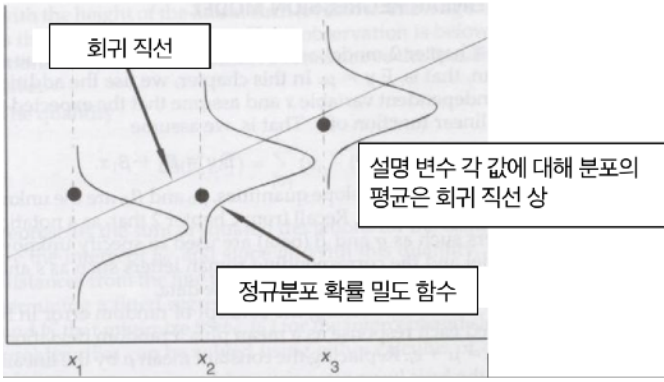
(b) 선형성 : 종속변수와 설명변수 간 함수관계는 선형

(c) 설명변수는 결정변수로 오차없이 측정할 수 있음 - 확률변수 아님

(d) 오차항 가정 $e_i \sim iidN(0,\sigma^2)$ 독립성, 정규성, 등분산성

(3) 오차항 가정 필요 이유

(a) 독립성(independent): 오차항은 서로 독립이다. 즉 각 오차는 서로 영향을 주지 않는다. 독립성 가정은 시계열 데이터(시간적 순서를 갖는 데이터) 경우에만 체크한다.



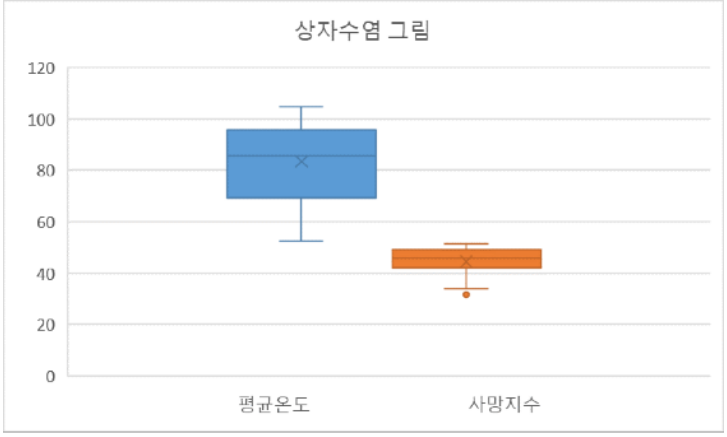
(b) 정규성(normality): 오차항은 정규 분포를 따른다. 이 가정은 F-검정 방법을 사용하기 위하여 반드시 필요하다. 오차항이 정규분포를 따르므로 종속변수도 정규분포를 따르고 다음이 성립 $y_i \sim iidN(\alpha + \beta x_i, \sigma^2)$

(c) 등분산성(Homoscedasticity): 오차항의 분산은 동일. 분산이 일정하다는 가정의 주어진 설명변수 값에서 관측되는 Y의 값의 분산이 일정하다는 의미와 같다. 분산이 다르면 설정된 회귀 모형이 적절함에도 불구하고 관측치가 직선에 모여 있지 않게 된다. 분산이 크므로 벗어나는 경향이 있다.

4. 데이터 정규변환

오차항 가정이 정규분포를 가정하므로 종속변수도 정규분포를 따르는 것이 회귀모형 추정의 결과의 신뢰성을 높인다. 이는 모든 데이터 분석에 적용된다. 하여, 데이터 분석 전 치우친 분포를 갖는 데이터에 대한 정규변환은 필수적이다.

자연형상, 사회현상으로부터 수집되는 대부분의 데이터는 좌우 대칭이거나 우로 치우친 형태를 갖는다(통계학 주요 확률분포함수 중 베타분포(수명이 상이한 두 부품의 수명 비)를 제외하고는 우로 치우친 형태, EXPonent(e)를 가지고 있음). 하여, 데이터 분석 전 우로 치우친 분포는 $X^{1/2} < X^{1/3} < Ln(X) < Log_{10}^{X}$ 정규변환을 실시하면 된다.



아래쪽 수염, 상자 길이가 다소 길지만 중위값과 평균이 유사하므로 좌우대칭으로 판단하는데 문제 없다. 물론 검정방법(Shapiro Wilks W-통계량, Anderson Darling AD-통계량)을 이용한 정규성 검정에 의해 검정되어야 한다.

5. 추정 ESTIMATION

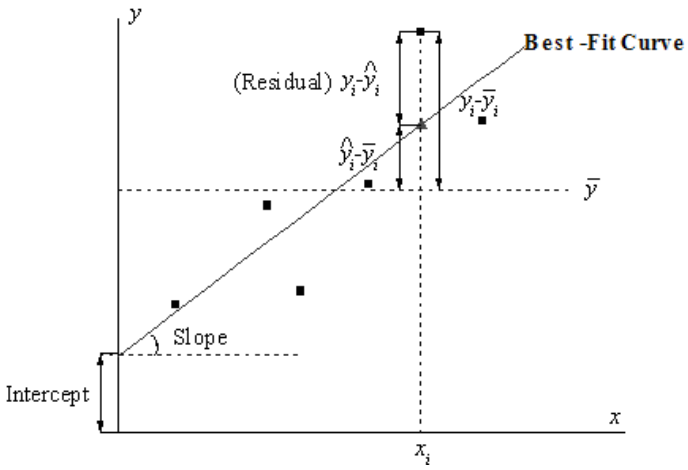
(1) 개념

- 데이터 (x_i, y_i) 활용하여 회귀계수 α, β 를 추정
- 추정된 회귀계수 $\hat{\alpha}, \hat{\beta}$ 를 이용하여 주어진 설명변수 값에서 종속변수 추정

  $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$ (fitted value 적합치)

(2) 최소자승법 Ordinary Least Square

- 관측점들을 가장 대표하는 직선 (best fit) 을 어떻게 구할 것인가?
- 관측점(y_i)과 직선($\alpha + \beta x_i$)의 거리를 최소화 하는 방법으로 구하자
- 거리를 수평선으로 하나? 수직선으로 하나? 수직선이 적절(y-축이 종속변수이고 종속변수 설명이므로)



$$\min_{\alpha,\beta} Q = \sum_i e_i^2 = \sum_i (y_i - \alpha - \beta x_i)^2$$

- 정규방정식

$$\frac{\partial Q}{\partial \alpha} = -2\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta} x_i) = 0$$

$$\frac{\partial Q}{\partial \beta} = -2\sum_{i=1}^{n} x_i (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0$$

- OLS 추정치

$$\hat{\beta} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \quad \left(\hat{\beta} = \frac{S_{xy}}{S_{xx}} \right), \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

6. 회귀계수 가설검정

(1) 개념

- 기울기 β : $H_0 : \beta = 0$ <=> 설명변수 X가 종속변수 설명 유의 않음 <=> 모형 유의하지 않음
- $H_0 : \beta = 0$ 귀무가설이 채택되면 $y_i = \alpha + e_i$ <- 종속변수를 설명하는 것은 종속변수 총 평균
- 절편 α에 대한 가설검정은 일반적으로 하지 않음 - (언제하나?) 원점을 통과하는가? 혹은 비용함수에서 절편은 고정비용이므로 이를 추정할 필요가 있을 경우

(2) 회귀계수 β 가설검정 - $H_0 : \beta = 0$ (설명변수 유의하지 않음)

(a) 귀무가설 $H_0 : \beta = \beta_0$ (설명변수 X의 유의성 검정 할 때 $H_0 : \beta = 0$)

(b) 대립가설 $H_0 : \beta \neq \beta_0$

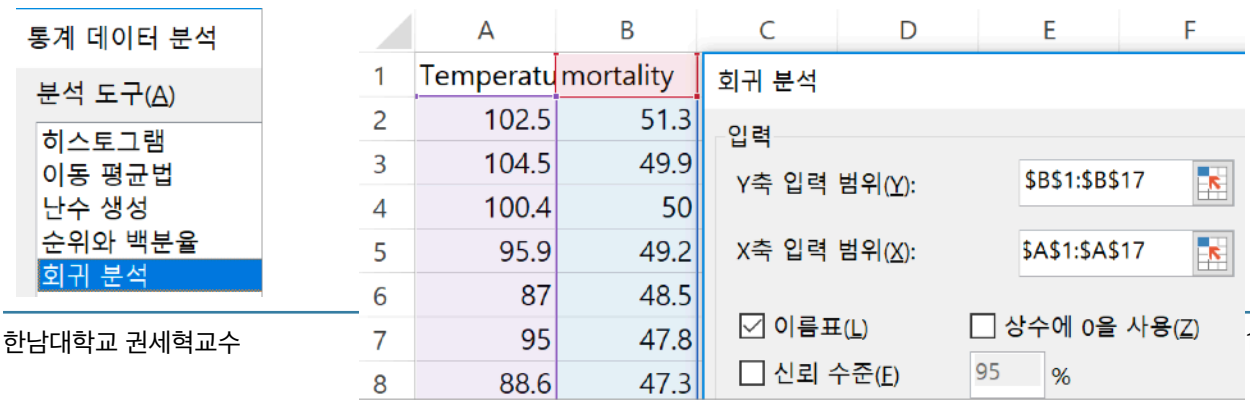
(c) OLS 추정량 (MVUE) : $\hat{\beta} = \frac{S_{XY}}{S_{XX}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$

(d) 추정량 $\hat{\beta}$의 샘플링 분포 $\frac{\hat{\beta} - \beta_0}{s(\hat{\beta})} \sim t(n-2)$, $s^2(\hat{\beta}) = \frac{\sigma^2}{S_{XX}}$

(e) 오차항의 분산 σ^2을 모르므로 추정치를 사용 : $\hat{\sigma^2} = MSE$ (분산분석 참고)

(f) $\hat{\beta}$의 분산($V(\hat{\beta})$) 추정치 : $s^2(\hat{\beta}) = \frac{MSE}{S_{XX}}$

사망지수(mortality index)=a+b*(평균온도, F: Fahrenheit) 회귀모형 추정하시오.



	계수	표준 오차	t 통계량	P-값
Y 절편	17.53816	4.064298	4.315175	0.000712
Temperature	0.324627	0.048037	6.757845	9.2E-06

$M = 17.5 + 0.32 * T$: 유의한 모형, 연평균 온도가 높아질수록 여성 암 사망지수는 높아진다. 온도가 1도 올라가면 사망지수는 0.32 높아진다. 산점도와 동일 결과이나 추정만으로는 이상치를 진단할 수 없다. - 잔차진단 필요

7. 잔차분석 RESIDUAL ANALYSIS

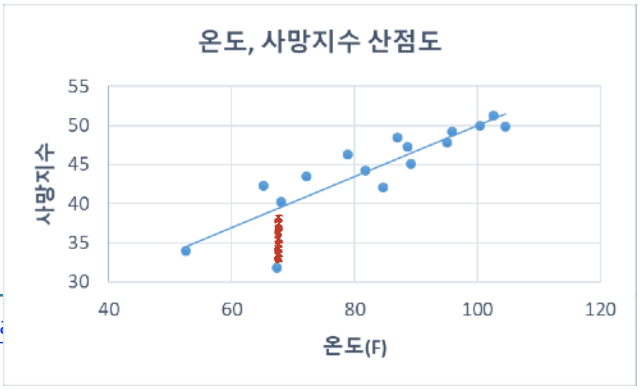
- 회귀분석은 설정된 회귀모형 $y_i = \alpha + \beta x_i + e_i$이 적합한지 (x_i, y_i) 쌍 관측 • 를 활용하여 회귀계수 β의 유의성을 설정하여 모형의 유의성을 판단
- 회귀분석 시작 시 오차항에 대한 가정과 함수의 선형성을 가정하였음 - 이 가정 하에 검정통계량의 샘플링분포(t-분포)도 구하였고 회귀계수에 대한 추론이 가능
- 오치항 가정이 성립되지 않으면 회귀분석 결과를 신뢰할 수 없음
- 하여, 회귀모형의 가정을 만족하는지 분석할 필요가 있음 - 이를 잔차분석이라 함
- 선형성 가정은 회귀모형 유의성 검정 결과와 동일하므로 잔차분석이라 함은 오차항에 대한 가정(정규성, 등분산성, 독립성)을 진단하고 문제를 해결하는 방법
- 독립성 진단은 시계열 자료에서만 검증함
- 추가적으로 회귀모형 추정결과에 영향을 주는 (이상,영향치) 진단까지 포함

(1) 잔차 계산

- 오차의 추정(MVUE) :

$$r_i = \hat{e}_i = y_i - \hat{y}_i = y_i - (\hat{\alpha} + \hat{\beta} x_i)$$

: (관측치-적합치)

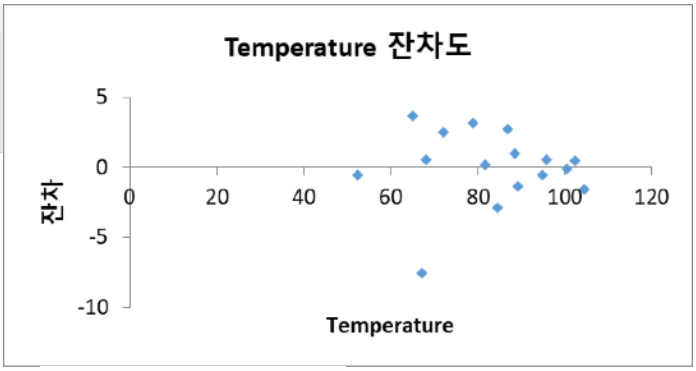


잔차, 종속변수 추정치=적합치 계산하고, 잔차 산점도를 그리시오.

잔차

☑ 잔차(R) ☑ 잔차도(D)

☑ 표준 잔차(T) ☑ 선적합도(I)



관측수	예측치 mortality	잔차	표준 잔차	
1	50.81237723	0.487623	0.180293	
2	51.46163031	-1.56163	-0.57739	
3	50.13066149	-0.13066	-0.04831	
4	48.66984204	0.530158	0.196019	
5	45.7806658	2.719334	1.005441	
6	48.37767815	-0.57768	-0.21359	
7	46.30006827	0.999932	0.369713	
8	46.4948442	-1.39484	-0.51573	
9	43.1511908	3.148809	1.164234	
10	45.0015621	-2.90156	-1.07282	
11	44.06014512	0.139855	0.05171	
12	40.97619295	2.523807	0.933147	
13	38.67134449	3.628656	1.341651	
14	39.64522413	0.554776	0.205122	
15	39.38552289	-7.58552	-2.80465	*
16	34.58105004	-0.58105	-0.21484	

=IF(ABS(D39)>2,"*","")

- 직선 위에 놓인 관측치의 잔차=0, 직선에 가까울수록 잔차 값은 적다.
- 오차의 정규분포 가정이 있으므로 잔차도 정규분포에 따른다.

(2) 잔차의 성질

- $E(r_i) = 0, s^2(r_i) = \sigma^2, \hat{s^2}(r_i) = MSE$
- $r_i \sim N(0, MSE)$

(3) 잔차의 종류

- 표준화 잔차 standardized residual

$$z_i = \frac{r_i}{\sqrt{MSE}} \sim N(0,1)$$ -

± 2 진단 - 이상치

	계수	표준 오차	t 통계량	P-값
Y 절편	21.55392456	2.785907	7.73677	3.22E-06
Temperature	0.283044647	0.032531	8.700634	8.82E-07

- 스튜던트 잔차 studentized residual $z_i = \frac{r_i}{\sqrt{MSE(1-h_{ii})}} \sim t(n-2),$

$h_{ii} = x_i'(X'X)^{-1}x_{i'}$ - ± 2 진단 - 이상치

잔차 산점도(엑셀은 종속변수 적합치 대신 설명변수가 X-축)에서 사전 진단한대로 15번째 관측치가 적합선(fitted line)으로 많이 벗어난 이상치이다. 관측치는 31.8, 예측치는 39.4로 잔차는 -7.59, 표준화 잔차는 -2.8이다. 설명변수가 관측된 범위 내에 있으므로 이상치로 진단되어 제외 후 추정하는 것이 적절하다.

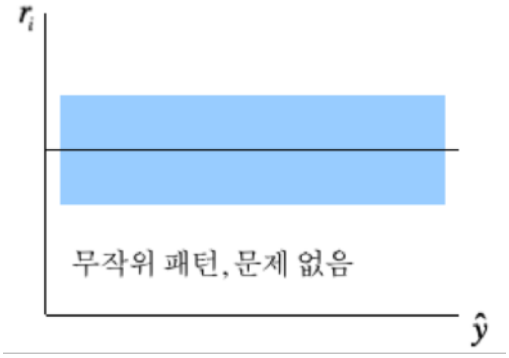
관측수	예측치 mortality	잔차	표준 잔차
1	50.56600084	0.733999	0.41906
2	51.13209014	-1.23209	-0.70343
3	49.97160708	0.028393	0.01621
4	48.69790617	0.502094	0.286659
5	46.17880882	2.321191	1.32523
6	48.44316599	-0.64317	-0.3672
7	46.63168025	0.66832	0.381562
8	46.80150704	-1.70151	-0.97144
9	43.88614718	2.413853	1.378133
10	45.49950167	-3.3995	-1.94087
11	44.67867219	-0.47867	-0.27329
12	41.98974805	1.510252	0.862243
13	39.98013106	2.319869	1.324475
14	40.829265	-0.62926	-0.35926
15	36.41376851	-2.41377	-1.37809

이상치 1개 문제 해결, 회귀계수의 유의확률이 작아져 설명변수 유의성 증가하였다. 모형의 설명력 척도인 결정계수도 증가한다.

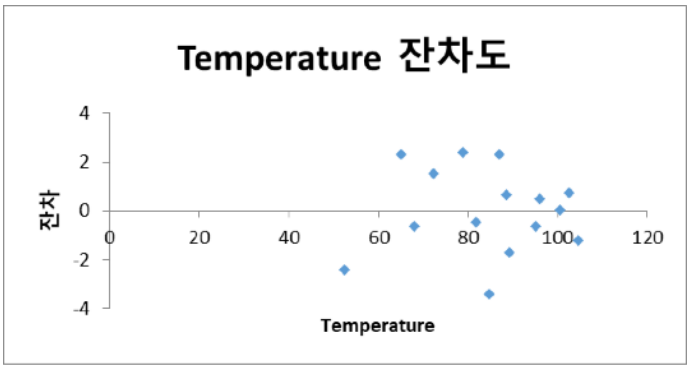
(4) 잔차진단 도구

(a) 잔차(r_i)와 종속변수 적합치($\hat{y}_i$) 산점도 : 단순회귀분석에서는 종속변수 적합치 대신 설명변수를 활용하면 된다.

잔차를 Y-축, 종속변수의 적합치를 X-축으로 하여 산점도를 그린다(잔차는 추정된 회귀 모형이 종속변수의 변동을 설명하지 못하는 부분에 해당하므로 산점도에 일정한 패턴이 있으면 안됨) 그리고 평균 0을 중심으로 무작위(random) 하게 흩어져 있어야 한다. 그리고 잔차가 크다는 것은 그 관측치가 이상치일 가능성 있다(또한 이 산점도에 의해 등분산성& 선형성도 진단한다)



잔차분석을 실시하고 분석결과를 정리하시오.

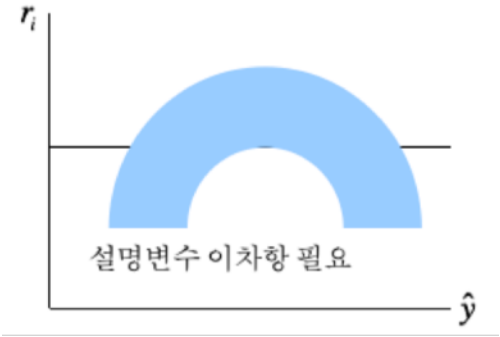


1개의 이상치를 제외하고는 표준화 잔차가 ± 2 초과하는 관측치가 없으므로 이상치는 더 이상 존재하지 않는다. 그리고 잔차 산점도에 특이 패턴이 발견되지 않으므로 최종 회귀분석 결과를 다음과 같이 정리하여 리포팅한다.

결정계수 = 85.3%				
변수	추정계수	표준 오차	t 통계량	P-값
절편	21.55392456	2.785907	7.73677	0.000
평균온도	0.283044647	0.032531	8.700634	0.000

$M = 21.5 + 0.28 * F(t = 8.7, p < 0.001)$: 평균온도와 사망지수는 양의 직선 관계가 존재하며, 평균온도가 1도 올라가면 사망지수는 0.28만큼 증가한다.

[이차항 문제 해결]

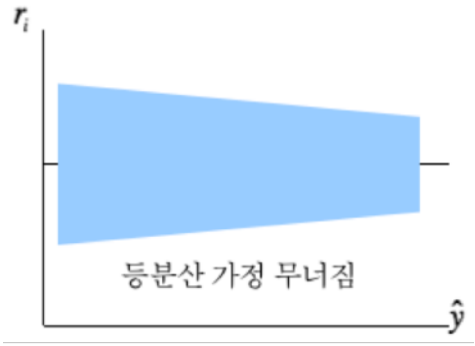


- 설명변수의 이차항(x_i^2)을 설명변수로 추가 : 그러나 일차항, 이차항이 동시에 있는 모형의 경우 다중공선성 문제 발생 가능성 높으므로 설명변수 ($x_i = \frac{(x_i - \bar{x})}{s(x)}$)를표준화하여 활용
- 방법 2 : 종속변수를 제곱근 변환($\sqrt{y_i}$)하여 회귀분석

사례 데이터 : 💾 ad.csv

기업의 광고액과 소비자 평가지수에 대한 데이터 회귀 분석하시오.

[등분산 문제 해결]



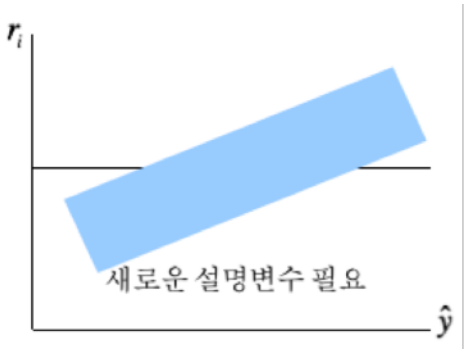
- 오른쪽 경우 팬 모양을 보이는데 이는 설명변수의 크기가 커짐에 따라 분산이 작아진다($V(e_i) \propto x_i$) 것을 의미하므로 등분산 가정이 무너진다.
- [해결방법 1] 회귀모형을 설명변수 x_i로 나눈 후 문제 해결 후 분석 - $\frac{y_i}{x_i} = \frac{\alpha}{x_i} + \beta + \frac{e_i}{x_i}$ 회귀모형
- [해결방법 2] WLS 가중최소자승법 $\min_{\alpha,\beta} \sum w_i(y_i - \alpha - \beta x_i)^2$, $w_i = \frac{1}{y_i^2}$

사례 데이터 : 💾 NFL.csv

NFL 선수의 드래프트 순위와 연봉을 조사한 데이터이다. 회귀분석을 실시하시오.

[새로운 설명변수 필요]

현재 모형에 고려된 설명변수만으로는 목표변수를 충분히 설명하지 못함 - 잔차에 현재 설명변수가 설명하지 못한 패턴이 존재함



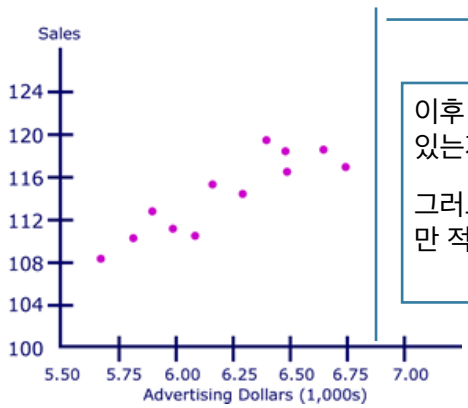
(b) 잔차(r_i)와 설명변수 (x_i) 산점도 : 설명변수가 2개 이상인 경우 (a)와 동일하게 활용 가능하다.

(c) 잔차(r_i)의 정규성 검정 : 오차항이 정규분포를 따른다는 가정 검증

8. 적합치 추정, 신뢰구간과 예측구간

(1) 적합치 fitted value : $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$

종속변수의 적합치를 구하거나 활용할 때 수집된 데이터의 설명변수 구간을 벗어난 영역에서는 하지 말아야 함



이후 설명변수 값에서는 (광고액과 판매량)이 직선의 관계가 있는지 어떤 근거도 없음

그러므로 수집된 데이터의 설명변수의 영역 (5.5, 6.75)에서만 적합모형을 활용하는 것이 적절함

(2) 신뢰구간 confidence interval for $E(y_0)$: 평균

- 설명변수 값이 x_0 로 주어졌을 때 종속변수 기대값 $E(y_0) = \alpha + \beta x_0$
- 점추정치 $\hat{E}(y_0) = \hat{\alpha} + \hat{\beta} x_0$

- 샘플링분포 $\frac{\hat{E(Y_0)} - E(Y_0)}{\hat{s\{E(Y_0)\}}} \sim t(n-2)$ $\hat{s\{E(Y_0)\}} = \sqrt{MSE[\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum(X_i - \overline{X})^2}]}$

(3) 예측구간 prediction interval for y_0 : 개별 관측값

- 새로운 설명변수 값이 x_0 로 주어졌을 때 종속변수 값 $y_{new} = \alpha + \beta x_0 + e_0$
- 점추정치 $\hat{y}_{new} = \hat{\alpha} + \hat{\beta} x_0$
- 샘플링분포 $\frac{\hat{Y}_{new} - E(\hat{Y}_{new})}{s\{\hat{Y}_{new}\}} \sim t(n-2)$ $s\{\hat{Y}_{new}\} = MSE[1 + \frac{1}{n} + \frac{(X_{new} - \overline{X})^2}{\sum(X_i - \overline{X})}]$

(4) comment

(a) 신뢰구간과 예측구간의 점 추정치는 $\hat{\alpha} + \hat{\beta} x_0$ 동일하지만 추정분산이 예측구간의 경우 MSE 만큼 넓어짐

(b) 적합치의 신뢰구간을 제시할 때는 구간 넓이는 좁은 “신뢰구간”을 사용

9. 분산분석적 접근 ANOVA APPROACH

(1) 개념

- 종속변수의 분산(변동)을 모형설명변동과 설명하지 못하는 변동으로 나누어 non-설명변동 대비 설명변동이 충분히 크면 모형이 유의하다고 판단
- 변동은 데이터 값의 변화에 대한 측정이므로 데이터의 정보와 동일함
- 정규분포를 따르는 확률변수(종속변수 y_i 가 이에 해당)의 변동(분산 계산과 동일)은 카이제곱 분포를 따르고 변동의 비 (카이제곱분포의 비)는 F-분포를 따르므로 이를 이용하여 모형의 유의성을 검증

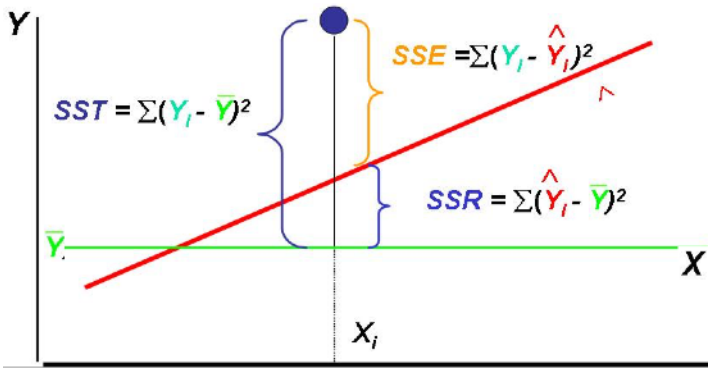
(2) 변동분할 variation decomposition

(a) 총변동 Total Sum of Square

$SST = \sum (y_i - \bar{y})^2$: 종속변수의 변동 (종속변수 값들의 변화)

(b) 모형변동 Regression(Model) SS

$SSR = \sum (\hat{y}_i - \bar{y})^2$: 종속변수 변동 중 설정된 모형에 의한 설명 부분



(c) 오차변동 Error SS $SSE = \sum (y_i - \hat{y}_i)^2$: 모형에 의해 설명되지 못하는 부분

(3) 변동의 분포

- $SSE = \sum (y_i - \hat{y}_i)^2 = \sum ((y_i - \bar{y}) - (\hat{y}_i - \bar{y}))^2$
- $\frac{\sum (y_i - \bar{y})^2}{\sigma^2} \sim \chi^2(n)$ since $y_i \sim iidN(\alpha + \beta x_i, \sigma^2)$
- $\frac{\sum (\hat{y}_i - \bar{y})^2}{\sigma^2} = \frac{\sum (\hat{\alpha} + \hat{\beta} x_i - (\alpha + \beta x_i))^2}{\sigma^2} \sim \chi^2(2)$ since $\hat{\alpha} \sim N(\alpha, ?), \hat{\beta} \sim N(\beta, ?)$
- $\frac{SSE}{\sigma^2} \sim \chi^2(n-2)$
- $SST = \frac{\sum (y_i - \bar{y}_i)^2}{\sigma^2} \sim \chi^2(n-1)$ since $y_i \sim iidN(\alpha + \beta x_i, \sigma^2)$

(4) 자유도 분할 : Cochran 정리

- 총변동의 자유도(관측치 중 자유로운 개수, 관측치 하나 하나는 독립적이고 정보를 갖고 있다)는 평균이 추정되었으므로 $(n-1)$ 이다.
- SSE의 자유도는 $(n-2)$ 이다. 왜냐하면 모수 (α, β) 두 개 추정되었기 때문이다.
- SSR의 자유도는 SST 자유도로부터 SSE 자유도를 뺀 값으로 1이다.

(5) 평균변동 Mean of (SSR, SSE) 및 기대평균변동 Expected MSE

(a) 평균회귀변동 MSR; Mean Sum of Squares for Regression

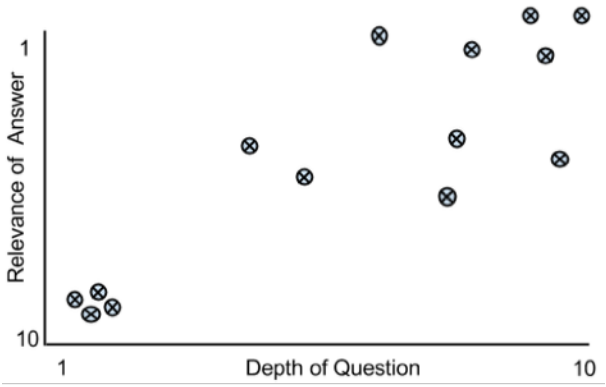
- $MSR = \frac{SSR}{df = 1}$ - 총변동 중 모형이 설명하는 변동 평균
- $E(MSR) = \sigma^2 + \beta^2 \sum (x_i - \bar{x})^2$ $\sigma^2 + \beta^2 \sum (x_i - \bar{x})^2$

(b) 평균오차변동 MSE; Mean Sum of Squares for Error

- $MSE = \frac{SSE}{df = n-2}$ - 총변동 중 모형이 설명하지 못하는 변동 평균
- 오차 분산 σ^2 의 추정치 => $E(MSE) = \sigma^2$

(6) F - 검정 - 회귀모형 $y_i = \alpha + \beta x_i$ 의 유의성 검정

- 검정통계량 $TS = \frac{MSR}{MSE} \sim F(1, n-2)$
- 검정통계량 평균 $E(TS) = \frac{\sigma^2 + \beta^2 \sum (x_i - \bar{x})^2}{\sigma^2} = 1 + \frac{\beta^2 \sum (x_i - \bar{x})^2}{\sigma^2}$
- 검정통계량이 1보다 커진다 <=> β가 0이 아니거나, 설명변수 관측값이 평균으로부터 많이 벗어나는 경우
- 검정통계량이 충분히 크다 <=> β가 0이 아님 ($H_0 : \beta = 0$ 기각) <=> 설정한 회귀모형은 유의



- 설명변수의 분산이 커지는 경우 β가 유의하지 않아도 F-값이 유의할 수 있음, 그리고 결정계수 (모형의 적합성 척도)도 커지게 되므로 반드시 산점도를 그려 설명변수 값의 범위에 주목할 필요가 있음

- 확률분포함수의 관계 : $F(1,n)=t^2(n)$ - 분산분석의 회귀모형 유의성($H_0:y_i\neq\alpha+\beta x_i$) 검정통계량의 샘플링분포 $F(1,n-2)$ = 회귀계수 유의성($H_0:\beta=0$) 검정통계량의 샘플링분포 $t(n-2)$의 제곱

(7) 분산분석표 ANOVA table

- 귀무가설 : 설정한 $y_i=\alpha+\beta x_i$가 유의하지 않음 <=> $H_0:\beta=0$ (설명변수 유의하지 않음)
- 대립가설 : 설정한 $y_i=\alpha+\beta x_i$가 유의

변동	자승합	자유도	평균자승합	기대평균	F-통계량
모형변동	SSR	1	MSR=SSR/1	$\sigma^2+\beta^2\sum(x_i-\bar{x})^2$	F=MSR/MSE ~F(1,n-2)
오차변동	SSE	n-2	MSE=SSE/(n-2)	$E(MSE)=\sigma^2$	
총변동	SST	n-1			

(8) 회귀모형 추정 결과 표

변수	추정계수	추정오차	t-통계량	유의확률
절편	$\hat{\alpha}$	$s(\hat{\alpha})$	$\hat{\alpha}/s(\hat{\alpha})$	
설명변수	$\hat{\beta}$	$s(\hat{\beta})$	$\hat{\beta}/s(\hat{\beta})$ $=\sqrt{F}$	분산분석 F-통계량 유의확률과 동일

10. 결정계수 DETERMINATION COEFFICIENT

(1) 정의

- $R^2=\frac{SSR}{SST}$: 총변동 중 회귀변동 비율 (%), $0<R^2<100(\%)$
- 종속변수 y_i와 적합치 $\hat{y}_i$의 상관계수 제곱 $r(Y,\hat{Y})=\sqrt{\frac{SSR}{SST}}$

(2) 해석

- 총변동 중 설정된 회귀모형이 설명하는 변동이므로 선택된 설명변수가 종속변수를 설명하는 능력을 비율로 나타낸 것임
- 결정계수가 낮은 경우 <=> 종속변수를 보다 잘 설명하는 다른 설명변수가 존재
- 결정계수가 80% (데이터 개수 30개 내외 수준) 이상이면 설정된 회귀모형이 종속변수를 충분히 설명하므로 더 이상 다른 설명변수를 찾을 필요는 없음 (회귀모형을 예측치 구하는 목적으로 사용하는 경우)

분산분석결과를 정리하시오.

분산 분석					
변동요인	자유도	제곱합	제곱 평균	F 비	유의한 F
회귀	1	250.11	250.11	75.7	0.0000
잔차	13	42.95	3.30		
총합	14	293.06			

회귀분석 통계량	
다중 상관계수	0.924
결정계수	0.853
조정된 결정계수	0.842
관측수	15

11. 상관계수와 관계

(1) 관계식

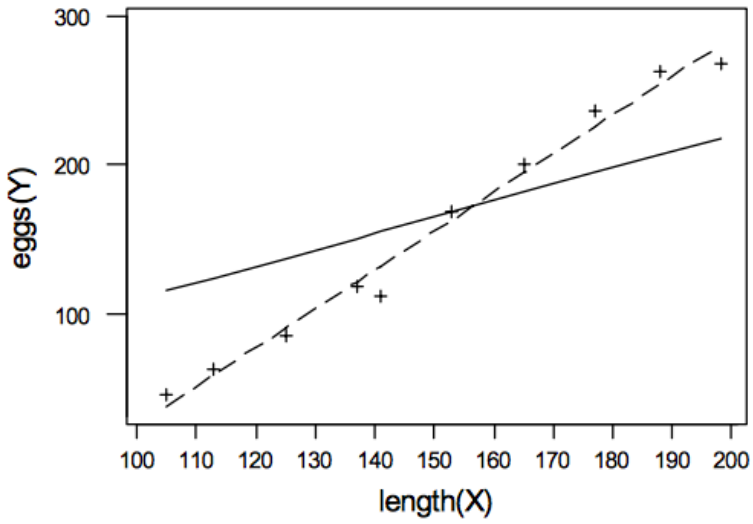
$$\hat{\beta} = \sqrt{\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}} \times r$$

$$r = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}} = \frac{E(X - E(X))E(Y - E(Y))}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$$

(2) 해석

- 상관계수와 회귀계수의 부호가 동일하며, 유의성 검정 시 동일하게 $t(n-2)$ 분포 사용

12. 원점을 지나는 회귀선

- 원점을 지나는 회귀 모형은 $y_i = \beta x_i + e_i$ 이므로 이 경우 회귀 계수 β 의 OLS는 $\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$ 이다. 일반 선형 회귀 모형 $y = \alpha + \beta x_i + e_i$ 에서 절편 (α) 이 0인 경우이다.
- 선형 회귀 모형에서 절편에 대한 가설 검정 $(H_0 : \alpha = 0)$ 을 실시하여 가설이 채택되면 원점을 지나는 회귀 직선을 사용하면 된다.
- 일반적으로 절편에 대해 관심이 없으므로(주로 기울기, 설명 변수의 영향) 절편에 대한 추정, 검정은 실시하지 않는다. 대신 분석하려는 상황(데이터)이 원점을 지나는 회귀 모형을 고려해야 한다면 처음부터 원점을 지나는 회귀 모형을 설정한다.
- 비용과 생산량과의 관계에서 고정비용을 고려할 경우에는 절편은 고정비용에 해당되므로 원점을 지나는 회귀모형이 아님
- 

사례분석 : KELLER "MANAGERIAL STATISTICS" 9TH VERSION

Case Study I : 💾 MLB_payroll.csv Keller 9th “Managerial Statistics”

메이저리그 30팀의 선수 총 연봉합과 승수를 조사한 데이터이다. 메이저 리그 팀을 운영하기 위한 선수 최소 연봉과 1승을 위하여 선수 연봉을 얼마나 지급해야 하는지 리포팅 하시오.

Case Study II : 💾 University.csv Keller 9th “Managerial Statistics”

캐나다 대학에서 우수 학생 선발을 위하여 고등학교 내신 성적 (최고 내신 6과목, 최고 내신 4과목+선형대수+영어 내신) 중 어느 것을 사용하여 신입생을 선발하는 것이 대학생활 우수 학생(학생의 우수성은 대학 학점 GPA로 측정)을 선발할 수 있는지 분석하시오.

Case Study III : 💾 Insurance.csv Keller 9th “Managerial Statistics”

H 도시에는 놀이공원과 박물관이 운영되고 있었다. 1991년부터 1995년까지 주 단위 관람객 수를 조사한 데이터이다. 33주~179주 사이에 박물관에 화재가 있어 운영이 중단되었다. 이 데이터를 활용하여 보험회사는 박물관에 얼마나 보상해 주어야 하는지 분석하시오.