

1장

서론

회귀분석

1.1

회귀분석이란 무엇인가?

회귀분석: 변수들 간의 함수적 관계를 찾는 방법을.

X Y

$$\begin{array}{ll}
 \text{설명변수} & \text{비율변수} \\
 \text{예측변수} & \text{통계변수} \\
 \text{독립변수} &
 \end{array}
 \Rightarrow Y = f(x_1, x_2, \dots, x_p) + \varepsilon_i \\
 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon_i \\
 \Rightarrow \text{선형회귀모형}, \varepsilon_i: \text{잔차}, \beta_0, \beta_1, \dots, \beta_p: \text{회귀계수}, \text{회귀정수}.$$

ex) (아들의 키) $\approx 33.73 + 0.516 \times (\text{아버지의 키})$

1.4

회귀분석의 단계

1. 문제에 대한 진술.

- 잘못 정의된 문제나 잘못 정식화된 질문은 노력은 낭비하게 된다.
- 주의 깊게 정식화 되지 않은 질문은 물어보기 않은 모형을 선택하게 된다.

ex) 고용주가 여자를 차별하고 있는지를 알아보자 한다.

① 평균적으로, 여자들이 동일한 평가를 받은 남자들보다 급여를 더 받고 있나?

X(예측변수): 업무평가, 성별, Y(반응변수): 급여

② 평균적으로, 여자들이 동일한 급여를 받은 남자들보다 좋은 평가를 받고 있나?

X(예측변수): 급여, 성별, Y(반응변수): 업무평가

2. 잡자적으로 적절한 변수들의 선택.

Y(반응변수): 단독주택의 가격

X(예측변수): 토지구역의 위치, 주택의 위치, 침실의 위치, 욕실의 개수 등.

3. 데이터 수집.

반응변수(Y) 예측변수 (x_1, x_2, \dots, x_p)

n	Y ₁ Y ₂ ⋮ Y _n	$\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$	$\Rightarrow n \times (p+1)$ 행렬	변수	양적
				질적	기사변수
					기변수

4. 모형 설정.

선형 ex) $Y = \beta_0 + \beta_1 x + \epsilon$

$$Y = \beta_0 + \beta_1 x^2 + \beta_2 \ln x + \epsilon$$

$$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon, (x^2 = x_1, \ln x = x_2)$$

→ 재표현 atau 변환

비선형 ex) $Y = \beta_0 + e^{\beta_1 x} + \epsilon$

- 양적반응변수 [한자 : 일변량
두개이상 : 다변량]

- 예측변수 [한자 : 도남
두개이상 : 다른]

모든 예측변수들이 질적변수 : 분산분석

예측변수들이 양적, 질적 둘다 있는 경우 : 긍부산 분석

빈용변수가 질적변수임 : 로지스틱 분포.

5. 적합방법

수집된 데이터 기초하여 모형의 모수를 추정 : 모수추정 & 모형적합.

가장 많이 사용되는 추정방법은 "최소제곱법"이다.

6. 모형 적합

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon$ 에서 데이터를 통해 회귀계수 ($\beta_0, \beta_1, \dots, \beta_p$)의 추정치 ($\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$)

를 구해 회귀방정식을 다음과 같이 쓰면

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p, \quad \hat{Y} \text{는 예측값이라고 한다}$$

1. 모형의 비판과 선택.

[필요한 가정들은 무엇인가?]

[이러한 가정들 각각에 대해서, 그 가정들이 타당한지를 어떻게 결정하는 것인가?]

[하나 이상의 가정들이 성립하지 않는 경우 무엇을 할 수 있는가?]

이러한 가정이 타당하지 않으면, 모형을 수정하고, 다시 가정이 타당하지 확인하는 과정을 거쳐, 만족하는 만한 결과가 얻어질 때까지 반복해야 한다.

- 표로 차트 ○ : 시작, 끝, ◇ : 예, 아니오로 나오는 질문, □ : 설명글

2. 회귀분석의 목적

회귀분석은 예측변수 $X (x_1, \dots, x_p)$ 과 빈용변수 Y 사이의 관계를 파악하고, 여러 목적으로 사용 가능하다.
(증가 예측 분석)

2장

단순선형회귀

2.1

소개

회귀모형의 형태 : $Y = f(x_1, x_2, \dots, x_p) + \epsilon$

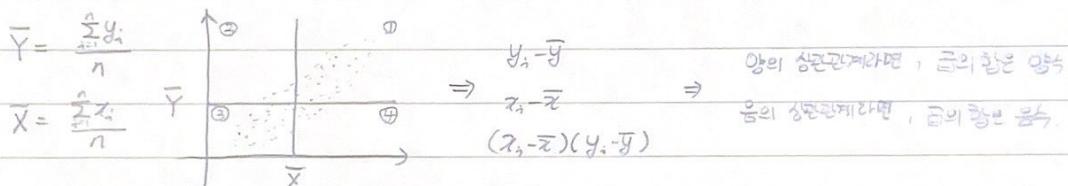
회귀모형 중에서 가장 단순한 형태는 $p=1$ 인 경우의 선형모형 : $Y = \beta_0 + \beta_1 x + \epsilon$

이때 이를 단순선형회귀라고 한다.

2.2

공분산과 상관계수

평균변수 \bar{Y} , 예측변수 X , n 개의 관찰자체 $\Rightarrow \{(X_i, Y_i)\}_{i=1}^n$



$$\text{공분산} : \text{Cov}(Y, X) = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{n-1}, \quad \text{Cov}(X, Y) > 0 : \text{양의 상관관계}, \quad \text{Cov}(X, Y) < 0 : \text{음의 상관관계}$$

공분산의 단점 (단위에 영향을 받는다)을 피하기 위해 Y, X 를 각각 표준화를 하여 공분산을 계산한다.

$$Y' = \frac{Y_i - \bar{Y}}{S_Y}, \quad X' = \frac{X_i - \bar{X}}{S_X} \quad (S_Y = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}}, \quad S_X = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}})$$

$$\text{상관계수} : \text{Cor}(Y, X) = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{Y_i - \bar{Y}}{S_Y} \right) \left(\frac{X_i - \bar{X}}{S_X} \right) = \frac{\text{Cov}(Y, X)}{S_Y \cdot S_X} = \frac{\sum (Y_i - \bar{Y})(X_i - \bar{X})}{\sqrt{\sum (Y_i - \bar{Y})^2} \sqrt{\sum (X_i - \bar{X})^2}}$$

Cor 은 척도(단위)에 불변한다. Y 와 X 사이의 선형관계의 강도를 나타낸다. ($-1 \leq \text{Cor}(Y, X) \leq 1$)

$\text{Cov}(Y, X) = \text{Cov}(X, Y)$, $\text{Cor}(Y, X) = \text{Cor}(X, Y)$ 로 대칭성을 가지고 있다.

단. 상관계수는 선형적 관계 특성만 알려주기 때문에 $\text{Cor}(Y, X) = 0$ 라는 것은 선형관계가 없음을

나타내지, 비선형적으로는 상관이 있을 수 있다.

2.3

실습.

2.4 단순선형회귀 모형

$$Y = \beta_0 + \beta_1 x + \varepsilon \quad : \text{단순선형회귀 모형}$$

β_0, β_1 : 회귀계수 or 회귀모수 (β_0 : 절편 β_1 : 기울기)

ε : 잔차 or 흔들오차

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i=1, 2, \dots, n \quad : \text{각 관측자에}$$

$\varepsilon_i \sim N(0, \sigma^2), \text{cov}(\varepsilon_i, \varepsilon_j) = 0, i \neq j$

$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

상관성 : 대형성을 가지기 때문에 X 와 Y 가 통통하게 중요하다.

회귀분석 : 선형적 특성에서 예측변수 X 보다 반응변수 Y 로 설명하기 풀기 때문에 Y 가 더 중요하다.

2.5 모수에 대한 추정

- 최소제곱법 사용.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$$

$$\varepsilon_i^2 = (y_i - \beta_0 - \beta_1 x_i)^2$$

ε_i^2 를 최소화 하기 위해 ε_i^2 를 최소화 하는 값을 구한다.

$$S(\beta_0, \beta_1) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\text{편미분사용} \Rightarrow \hat{\beta}_0 = \bar{y}, \hat{\beta}_1 = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

(증명필요)

즉. β_0, β_1 의 추정치 $\hat{\beta}_0, \hat{\beta}_1$ 를 대입하여 모델곡선을 최소제곱회귀선이라고 한다.

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X \quad : \text{최소제곱회귀선.}$$

- 통계학적으로 $\hat{\beta}_0, \hat{\beta}_1$ 는 β_0, β_1 의 불편추정량 값이다 \Rightarrow (증명필요)

직접값 (\hat{y}_i) = $\hat{\beta}_0 + \hat{\beta}_1 x_i$ 에서 실제값 (y_i)과의 차이를 진차 or 최소제곱진차라고 한다.

$$\text{진차} (\varepsilon_i) = y_i - \hat{y}_i, \quad \sum \varepsilon_i = 0 \text{ 포함 성질로 만족한다.}$$

2.6 개선검정

$\hat{\beta}_0$ 와 $\hat{\beta}_1$ 의 불신은 다음과 같다.

$$\text{Var}(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right] \quad \text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \Rightarrow \text{(증명필요)}$$

회귀계수 $\hat{\beta}_0$ 와 $\hat{\beta}_1$ 외의 진차의 분산인 σ^2 역시 추정할 필요가 있다.

$$\hat{\sigma}^2 = \frac{\sum \varepsilon_i^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{\sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2} = \frac{\text{SSE}}{n-2} \quad (n-2 \text{는 } n \text{개의 } y_i \text{에서 회귀계수의 개수 } 2 \text{를 뺀 값으로 SSE의 자유도이기도 함이나 })$$

SSE는 진차 ε_i 의 제곱합이고, 이렇게 정의된 $\hat{\sigma}^2$ 는 σ^2 의 불편추정량이 된다.

(증명필요)

즉, $\hat{\beta}_0$ 와 $\hat{\beta}_1$ 의 분포는 다음과 같다.

$$\hat{\beta}_0 \sim N(\beta_0, \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2} \right] \sigma^2), \quad \hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum(x_i - \bar{x})^2})$$

또한 σ^2 대신 SSE 를 넣고, 표준편차의 추정치인 표준오차를 구해보면.

$$S.E.(\hat{\beta}_0) = \sqrt{\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2} \right]} = \sqrt{\frac{SSE}{n-2} \cdot \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2} \right]}$$

$$S.E.(\hat{\beta}_1) = \sqrt{\frac{1}{\sum(x_i - \bar{x})^2}} = \sqrt{\frac{SSE}{(n-2) \times \sum(x_i - \bar{x})^2}}$$

표준오차가 작을수록 추정량의 정밀도가 높다고 할수 있다.

* 보통 변수 Y에 예측변수 X가 영향을 미치는지 β_0 로 평가를 하려한다

$$H_0: \beta_1 = 0 \quad vs \quad H_1: \beta_1 \neq 0 \quad (\text{양측})$$

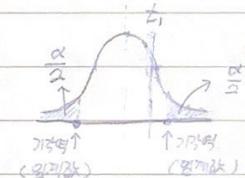
일단 H_0 가 맞다고 가정하고 틀고 오류가 생기면 기각하는 방법을 사용한다.

$$\hat{\beta}_1 \sim N(0, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}) \Rightarrow \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{\sum(x_i - \bar{x})^2}}} \sim N(0, 1) \quad \text{그 계산하고 결과를 모수에서 본가하므로}$$

$$\hat{\beta}_1 \sim N(0, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}) \Rightarrow \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{\sum(x_i - \bar{x})^2}}} \sim t(n-2), \quad \text{표본인 표준오차를 사용해 계산한다.}$$

① 즉 위에 검정에 대한 검정통계량은 다음과 같다.

$$t_1 = \frac{\hat{\beta}_1}{S.E.(\hat{\beta}_1)}, \quad H_0 \text{ 를 기각함 } P(|t_1|) \leq \alpha, \quad H_0 \text{ 를 기각안함, } P(|t_1|) > \alpha$$



② 티다는 통계량으로는 P-value 가 있다. P-value 란 t分布를 따르는 T의 절댓값이 n의 절댓값보다 큼값을 가질 확률이다.

$$H_0 \text{ 를 기각함. } P(|t_1|) \leq \alpha \\ H_0 \text{ 를 기각안함 } P(|t_1|) > \alpha$$



* β_0 가 0이 아닌 β_0^* (특정상수)에 대하여 검정을 진행하면 다음과 같다

$$H_0: \beta_1 = \beta_1^* \quad vs \quad H_1: \beta_1 \neq \beta_1^*$$

$$t_1 = \frac{\hat{\beta}_1 - \beta_1^*}{S.E.(\hat{\beta}_1)} \quad H_0 \text{ 를 기각함. } |t_1| \geq t_{n-2, \alpha/2} \\ H_0 \text{ 를 기각안함 } |t_1| < t_{n-2, \alpha/2}$$

* β_1 가 아닌 β_0 에 대한 가설검정도 가능하다. 특정상수 β_0^* 에 걸맞을 하면 다음과 같다.

$$H_0: \beta_0 = \beta_0^* \quad \text{Vs} \quad H_1: \beta_0 \neq \beta_0^*$$

$$t_0 = \frac{\hat{\beta}_0 - \beta_0^*}{\text{S.e.}(\hat{\beta}_0)} \sim t(n-2)$$

구하기엔 H_0 가 맞다는 가정하에 t_0 은 자유도 $n-2$ 인 t 분포를 따른다.

H_0 를 기각함. $|t_0| \geq t(n-2)$

H_0 를 기각 안함 $|t_0| < t(n-2)$

* β_1 가 아닌 상관계수 P 로 선형관계를 검정하는 방법은 다음과 같다.

$$H_0: P=0 \quad \text{Vs} \quad H_1: P \neq 0$$

$$t_1 = \frac{\text{Cor}(X, Y) \cdot \sqrt{n-2}}{\sqrt{1 - \text{Cor}(X, Y)^2}} \sim t(n-2)$$

구하기엔 H_0 가 맞다는 가정하에 t_1 은 $t(n-2)$ 를 따르므로 다음과 같다

H_0 를 기각함. $|t_1| \geq t(n-2)$

H_0 를 기각 안함 $|t_1| < t(n-2)$

이를 순서대로 (추정값, 표준오차, ±통계량, P값) 정리해 놓은 것이 계수표이다.

변수	계수	표준오차	±통계량	P값
절편	$\hat{\beta}_0$	S.e. ($\hat{\beta}_0$)	t_0	P_0
X	$\hat{\beta}_1$	S.e. ($\hat{\beta}_1$)	t_1	P_1

* P값이 α 보다 작으면 기각

2.1 선형구간

$$\beta_0 \in \left[\hat{\beta}_0 \pm t_{n-2, \frac{\alpha}{2}} \times \text{S.e.}(\hat{\beta}_0) \right]$$

$$\beta_1 \in \left[\hat{\beta}_1 \pm t_{n-2, \frac{\alpha}{2}} \times \text{S.e.}(\hat{\beta}_1) \right]$$

2. 8

예측

① 예측변수 $X = x_0$ 에 대응하는 비용변수 Y 의 값의 대한 예측

$$\text{즉 } \hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \varepsilon_0, \quad \varepsilon_0 \sim N(0, \sigma^2) \text{ 의 예측값인}$$

$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ 이라고 할수있다, 이때 \hat{y}_0 에 대한 신뢰구간은 다음과 같다. 그리고 이를 예측한계라고 한다.

$$y_0 \in \left\{ \hat{y}_0 \pm t_{n-2, \alpha/2} \times S.e.(\hat{y}_0) \right\} \Rightarrow \text{예측한계}$$

$$(\text{또, } S.e.(\hat{y}_0) = \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}} \text{ 이다.}) \Rightarrow \text{증명 필요}$$

② 예측변수 $X = x_0$ 에 대응하는 평균비용 μ_0 에 대한 예측

$$\text{즉 } \mu_0 = \beta_0 + \beta_1 x_0 \text{ 의 예측값인 } \hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \text{ 라고 할수있다.}$$

이때 $\hat{\mu}_0$ 에 대한 신뢰구간은 다음과 같다. 그리고 이것을 신뢰한계라고 한다

$$\mu_0 \in \left\{ \hat{\mu}_0 \pm t_{n-2, \alpha/2} \times S.e.(\hat{\mu}_0) \right\} \text{ 단 } S.e.(\hat{\mu}_0) = \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}} \Rightarrow \text{증명 필요}$$

예측한계와 신뢰한계를 비교했을 때 하나의 개체 예측이 평균여대한 예측보다 불안정 하므로 신뢰한계가

예측한계가 더 크게 당면하다. 시행 n 이 무수히 많을 때 다음이 성립된다. 다만 절댓값은 같다.

$$n \rightarrow \infty \quad S.e.(\hat{y}_0) \rightarrow 0$$

$$S.e.(\hat{\mu}_0) \rightarrow 0 \quad \text{이다.}$$

2. 9

적합성의 측정

적합성 : 선형모형을 적합 후, 자료의 모형이 얼마나 잘 적합한지 확인하는 것

계수표를 통해 Y 와 X 의 선형관계의 "강도"를 측정가능하다. t 값이 클수록, P 값이 0에 가까울수록 선형관계가

크다고 할수있다. 또는 선점도가 직선에 가까울수록, $Cor(X, Y)$ 가 1에 가까울수록 선형관계가 크다.

Y 와 \hat{Y} 의 선점도를 그래서 일정선에 가깝거나 $Cor(Y, \hat{Y}) = |Cor(Y, X)|$ 가 1에 가까울수록 선형관계가 크다.

자료가 없을때는 \hat{Y} 가 가장 좋은 측정치이지만 자료가 있을때는 \hat{Y} 가 가장 좋은 측정치이다

또한 적합도지수 or 결정계수라고 불리는 R^2 을 구할수 있는데, $0 \leq R^2 \leq 1$ 의 범위 가지고 1에 가까울수록 선율성이 좋다

$$SST (\text{평균으로부터의 평차의 제곱합}) = SSR (\text{적합에 기인한 평차의 제곱합}) + SSE (\text{잔차의 제곱합})$$

$$\sum(Y_i - \bar{Y})^2 = \sum(\hat{Y}_i - \bar{Y})^2 + \sum(Y_i - \hat{Y}_i)^2$$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = [Cor(Y, X)]^2 = [Cor(Y, \hat{Y})]^2$$

2.10 율점을 통과하는 회귀선

즉 절편 β_0 가 없는 회귀선으로 기본모형은 다음과 같다.

$$Y_i = \beta_1 X_i + \varepsilon_i \quad (i=1, 2, \dots, n) \quad \varepsilon_i \sim N(0, \sigma^2) \quad Y_i \sim N(\beta_1 X_i, \sigma^2)$$

이를 실현함이 없는 모형이라 한다.

β_1 을 최소제곱법을 이용하여 결정치를 구하면

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} \quad \text{이다.} \quad \Rightarrow (\text{증명필요})$$

또한 표준오차 표현

$$S.e(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum x_i^2}}, \quad \hat{\sigma} = \sqrt{\frac{\sum e_i^2}{n-1}} = \sqrt{\frac{SSE}{n-1}} \quad \Rightarrow (\text{증명필요})$$

이렇게 구해진다. 그리고 결정계수 R^2 의 관계식은 다음과 같다

$$SST = SSR + SSE = \sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2 \quad (\text{평균을 제거})$$

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

2.11 사소한 회귀모형

즉 예측변수를 가지지 않는 모형이므로 기본모형은 다음과 같다.

$$Y_i = \beta_0 + \varepsilon_i \quad (i=1, 2, \dots, n) \quad \varepsilon_i \sim N(0, \sigma^2) \quad Y_i \sim N(\beta_0, \sigma^2)$$

β_0 를 최소제곱법을 이용해 미분값이 0 이 되도록 하여 β_0 의 결정치를 구해보면 다음과 같다.

$$\hat{\beta}_0 = \bar{y}$$

또한 $\hat{\beta}_0$ 의 분산을 유도해보면 다음과 같다.

$$\text{Var}(\hat{\beta}_0) = \frac{\sigma^2}{n}$$

β_0 의 최소제곱 결정치는 다음과 같다.

$$\hat{\beta}_0 = \bar{y}$$

따라서 적합한 \hat{y}_i 및 진화 e_i 는 다음과 같다.

$$\hat{y}_i = \bar{y} \quad e_i = y_i - \bar{y}$$

이때 고려되는 가설은 $H_0: \beta_0 = 0$ vs $H_1: \beta_0 \neq 0$ 이고 이때 검정통계량은 다음은 t 통계량이다.

이를 일표본 t 검정이라고 한다.

$$t = \frac{\bar{y} - 0}{S.e(\bar{y})} = \frac{\bar{y}}{S_y / \sqrt{n}}, \quad S_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}$$

또 다른 사소한 회귀모형으로 "대응 일표본 t 검정" 이라 한다. 예로 Y_1 시장판매 총량, Y_2 공급판매 총량, $Y = Y_1 - Y_2$ 라고 하자.

사실 Y_1, Y_2 이 변수라면 $Y = Y_1 - Y_2$ 를, 일표본으로 생각할 수 있다. $Y = \mu + \varepsilon$ 의 모형으로 가정, $H_0: \mu = 0$ vs $H_1: \mu > 0$ 으로 검정 가능하다.

3장

다중선형회귀

3.1

소개

3.2

데이터와 모형에 대한 서술

Y 와 p 개의 X 에 대한 n 개 관찰자로이 선형으로 표현가능하다.

이를 다중선형회귀라고 한다.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon, \quad \beta_0, \beta_1, \dots, \beta_p : \text{회귀계수 or 브카계수}$$

ε : 특정오차 or 잔류오차 (이제한경우 \times)

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} + \varepsilon_i \quad (i=1, 2, \dots, n), \quad \varepsilon_i \sim N(0, \sigma^2)$$

$$Y_i \sim N(\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}, \sigma^2), \quad \text{Cov}(\varepsilon_i, \varepsilon_j) = 0, (i \neq j)$$

• 다중선형회귀는 단순선형회귀의 일반화된 모형이다 $p=1$ 인 경우 단순회귀로 쓰여진다.

3.3

사례 :

	령	Y	X_1	X_2	
1	43	(51)	.	.	$\rightarrow 35\text{명} \pm 18\text{명} \text{ 평균}$
2	$= \frac{18}{35} \times 100 = 51.42 \approx 51$

3.4

모수추정

단순선형모형과 마찬가지로 최소제곱법을 통해 모수를 추정한다.

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots +$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \dots + \hat{\beta}_p X_{ip}$$

$$S(\beta_0, \beta_1, \dots, \beta_p) = \sum (Y_i - \beta_0 - \beta_1 X_{i1} - \dots - \beta_p X_{ip})^2$$

• 절규방정식 ($\frac{\partial S}{\partial \beta_0} = \frac{\partial S}{\partial \beta_1} = \dots = \frac{\partial S}{\partial \beta_p} = 0$)를 통해서 추정치 $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ 를 구한다

단. $n-1$ 개의 방정식을 연립하여 험들이기 때문에 행렬을 만들고 역행렬을 만들어 구한다.

추정치 $(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$ 를 구한다면 적합도를 구해보면 다음과 같다

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \dots + \hat{\beta}_p X_{ip} \quad i=1, 2, \dots, n$$

잔차와 σ^2 의 불편성장을 구해보면 다음과 같다

$$e_i = y_i - \hat{y}_i \quad i=1, 2, \dots, n$$

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-(p+1)} = \frac{\sum (y_i - \hat{y}_i)^2}{n-(p+1)} = \frac{SSE}{n-p-1}, \quad \text{이때 } n-(p+1) \text{를 자유도라고 한다.}$$

3.5 회귀계수에 대한 해석

다중회귀 분석에서 $p=1$: 단순회귀분석

$p=2$: 평면

$p \geq 3$: 초평면 으로 표현하게 된다.

상수항 계수 β_0 는 $X_1 = X_2 = \dots = X_p = 0$ 일 때 Y 의 기댓값이다.

회귀계수 β_i 는 X_j 를 제외한 나머지 X 를 상수 취급하고, X_j 가 한 단위 증가할 때 Y 의 증가량으로 해석 가능

3.3절 예제 데이터를 가지고 $p=2$ 인 다중회귀 분석을 해보자.

$$\hat{Y} = 15.3216 + 0.7803X_1 - 0.0502X_2 \\ = \beta_0 + \beta_1 \cdot X_1 - \beta_2 \cdot X_2$$

X_2 에 대하여 조정된 후에 얻어진 X_1 의 효과 : $\beta_1 = 0.7803$

X_1 에 대하여 조정된 후에 얻어진 X_2 의 효과 : $\beta_2 = -0.0502$

① X_1 에 Y 를 접합한 모형이 다음과 같고, 이때 잔차를 e_{Y,X_1} 라고 하자

$$\hat{Y} = 14.3163 + 0.74610 X_1$$

② X_1 에 X_2 를 접합한 모형이 다음과 같고, 이때 잔차를 e_{X_2,X_1} 라고 하자

$$\hat{X}_2 = 18.9654 + 0.513032X_1$$

③ 보통변수가 e_{Y,X_1} 이고 예측변수는 e_{X_2,X_1} 라고 할 때 이를 추정을 했으면 다음과 같다.

$$\hat{e}_{Y,X_1} = 0 - 0.0502 \hat{e}_{X_2,X_1}$$

즉 Y, X_1, X_2 의 관계에서 X_1 에 대한 영향은 ①, ②에서 없었고 ③에서 Y, X_2 에 대한 영향을 했다.

X_1 에 대하여 조정된 후의 얻어진 X_2 의 효과 만이 남아 있게 된다.

여러 변수들이 서로 무상관이 아니라면 어떤 설명변수에 대한 단순회귀와 다중회귀계수는 근본적으로 서로 다르다.

3.6

중심화와 척도화.

회귀계수는 단위의 영향을 받는다. 이를 해결하기 위해 중심화와 척도화를 진행한다.

- 1) 절편이 있는 일반적인 다중회귀모형에 대한 중심화와 척도화를 알아보자.

중심화 \Rightarrow 각 관측값에서 모든 관측값의 평균을 빼는 것으로 일어난다.

$$y_i \Rightarrow y_i - \bar{y} \quad i=1, 2, \dots, n$$

$$x_{ij} \Rightarrow x_{ij} - \bar{x}_j \quad i=1, 2, \dots, n \quad j=1, 2, \dots, p$$

중심화된 자료들의 표준은 0이 되고, 척도화가 가능하다. 척도화에는 단위간이 척도화와 표준화가 많이 사용된다

① 단위길이 척도화

$$y_i - \bar{y} \Rightarrow \frac{y_i - \bar{y}}{s_y} = \tilde{y}_i, \quad L_y = \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad i=1, 2, \dots, n$$

$$x_{ij} - \bar{x}_j \Rightarrow \frac{x_{ij} - \bar{x}_j}{s_j} = \tilde{x}_{ij}, \quad L_j = \sqrt{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}, \quad j=1, 2, \dots, p, \quad i=1, 2, \dots, n$$

L_y, L_j 는 중심화변수 $y_i - \bar{y}$ 와 $x_{ij} - \bar{x}_j$ 의 길이가 된다. 따라서 \tilde{y}_i 와 \tilde{x}_{ij} 는 평균 0, 길이 1을 가지게 된다.

* 원변수 x_j 와 x_k 의 대하는 상관계수는 단위길이 척도화된 변수를 사용하면 간편히 계산 가능하다.

$$\text{Cor}(x_j, x_k) = \frac{1}{n} \sum_{i=1}^n \tilde{x}_{ij} \tilde{x}_{ik}$$

② 표준화

$$y_i - \bar{y} \Rightarrow \frac{y_i - \bar{y}}{s_y} = \tilde{y}_i, \quad S_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}, \quad i=1, 2, \dots, n$$

$$x_{ij} - \bar{x}_j \Rightarrow \frac{x_{ij} - \bar{x}_j}{s_j} = \tilde{x}_{ij}, \quad S_j = \sqrt{\frac{\sum (x_{ij} - \bar{x}_j)^2}{n-1}}, \quad i=1, 2, \dots, n, \quad j=1, 2, \dots, p$$

S_y, S_j 는 y, x_j 의 표준편차를 의미하고 $\tilde{y}, \tilde{x}_{ij}$ 는 평균 0, 표준편차 1을 갖는다

상관관계를 계산할 때 단위길이 척도화나 표준화 모두 편리하게 이용 가능하다.

- 2) 절편이 없는 다중회귀모형에 대한 중심화와 척도화.

이 모형의 경우 중심화가 따로 없기 때문에 바로 척도화가 가능하다.

① 단위길이 척도화

$$y_i \Rightarrow \frac{y_i}{L_y} = \tilde{y}_i, \quad L_y = \sqrt{\sum_{i=1}^n y_i^2}, \quad i=1, 2, \dots, n$$

$$x_{ij} \Rightarrow \frac{x_{ij}}{L_j} = \tilde{x}_{ij}, \quad L_j = \sqrt{\sum_{i=1}^n x_{ij}^2}, \quad i=1, 2, \dots, n, \quad j=1, 2, \dots, p$$

중심과 척도화를 통하여 변환된 변수는 원 변수로 되돌리 가능하다.

$$\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p \rightarrow \hat{\theta}_0, \hat{\theta}_1, \dots, \hat{\theta}_p$$

① 중복화인 경우, $\hat{\beta}_0$ 를 제외한 $\hat{\beta}_1, \dots, \hat{\beta}_p$ 는 변함이 없다. 즉 $\hat{\beta}_0 \rightarrow \hat{\theta}_0$ 만이 변한다.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2 - \dots - \hat{\beta}_p x_p, \quad \hat{\theta}_0 = 0$$

$$\hat{\beta}_i = \hat{\theta}_i \quad (i=1, 2, \dots, p)$$

② 척도화인 경우, 치환하면 같은 원리를 있으니 증명하여야 한다.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2 - \dots - \hat{\beta}_p x_p, \quad \hat{\theta}_0 = 0$$

$$\hat{\beta}_i = \frac{s_y}{s_x} \hat{\theta}_i, \quad (i=1, 2, \dots, p)$$

3.7 최소제곱추정량의 성질

추정량 $\hat{\beta}_j$ ($j=0, 1, \dots, p$)는 β_j 의 불편추정량이며 분산은 $\nabla^2 C_{jj}$ 이다.

그리고 $\hat{\beta}_j$ 와 $\hat{\beta}_k$ 의 공분산은 $\nabla^2 C_{jk}$ ($j \neq k$) 가된다. 여기서 C_{jj} 와 C_{jk} 는 "수정된 제곱차합" 행렬이다.

이렇게 정의된 $\hat{\beta}_j$ 들은 모두 불편추정량 중에서 최소분산을 가지며 BLUE라 한다.

$$W = \frac{SSE}{\sigma^2} \sim \chi^2(n-(p+1))$$

$\hat{\beta}_j$ 는 평균이 $\beta = (\beta_0, \beta_1, \dots, \beta_p)$ 이고 분산공분산행렬은 수정된 제곱차합 행렬인 다변량 정규분포를 따른다.

$$\hat{\beta} \sim MVN(\beta, C), \quad C = [C_{ij}]$$

3.8 다중상관계수

Y 와 X_1, \dots, X_p 사이의 선형 정도는 어떤가?

① Y 와 \hat{Y} 의 선형

$$\text{② } Y \text{와 } \hat{Y} \text{의 상관계수}, \quad Cor(Y, \hat{Y}) = \frac{\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}}$$

$$\text{③ 결정계수 } R^2 = [Cor(Y, \hat{Y})]^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}, \quad 0 \leq R^2 \leq 1, \quad (0이거나 1이면 선형관례)$$

R 은 ($= \sqrt{R^2}$) 다중상관계수라고 불리기도 한다.

단, 예측변수의 수, p 를 증가시키면 R^2 이 무작위 키기는 단점이 있지만, 수정된 결정계수를 통해 단점을 없앤다.

$$\text{④ 수정된 결정계수} \quad R_a^2 = 1 - \frac{\frac{SSE}{n-(p+1)}}{\frac{SST}{n-1}}$$

3.9

개별 회귀계수들에 대한 흐름

특정 회귀계수가 유의한지에 대한 가설 검정을 진행해보자 (특정상수 β_j^0 , 05가능하다)

$$H_0: \beta_j = \beta_j^0 \quad vs \quad H_1: \beta_j \neq \beta_j^0 \quad (j=1, 2, \dots, p \text{ 중 } 1개)$$

검정통계량 : $t_j = \frac{\hat{\beta}_j - \beta_j^0}{\text{S.e.}(\hat{\beta}_j)} \sim t(n-(p+1))$

귀무가설 H_0 가 맞다는 가정하에 t_j 는 자유도가 $n-(p+1)$ 인 t 분포를 따른다.

H_0 를 기각 : $|t_j| \geq t_{n-(p+1), \alpha/2}$

H_0 를 채택 : $|t_j| < t_{n-(p+1), \alpha/2}$

위 검정에서 H_0 를 기각한다는 것은 β_j 가 비통제변수 X_j 에 대해 유의하다라는 결과를 낼 수 있다.

신뢰구간은 다음과 같다.

$$\beta_j \in \left\{ \hat{\beta}_j \pm t_{n-(p+1), \alpha/2} \cdot \text{S.e.}(\hat{\beta}_j) \right\}$$

3.10

선형모형에서의 가설검정

원래의 다중선형회귀모형은 원진모형 (FM)으로 놓고, 검정하거나 하는 상황을 단축하는 모형은 확소모형 (RM)으로 두고

가설을 검정하는 것이다. 확소모형은 특수한 경우로 원진모형이 내포되어 있어야 한다.

이때 우리는 다음의 가설을 진행한다.

H_0 : 확소모형 (RM)이 적절하다 vs H_1 : not H_0 , 원진모형이 적절하다.

\hat{y}_i 와 \hat{y}_i^* 를 각각 FM과 RM의 적합값이라고 했을 때,

$$\text{원진모형의 비적합도} : SSE(FM) = \sum (y_i - \hat{y}_i)^2$$

$$\text{확소모형의 비적합도} : SSE(RM) = \sum (y_i - \hat{y}_i^*)^2$$

라고 할 수 있으며, 모두가 원진모형에 비슷하거나 더 많이 쓰이므로 항상 $SSE(FM) \leq SSE(RM)$ 이 성립한다.

$SSE(RM) - SSE(FM)$ 의 값이 크면 원진모형이 적절, 값이 작으면 확소모형이 적절이라고 볼 수 있다.

위의 검정에서 검정통계량은 다음과 같다. (원진모형: $p+1$ 개, 확소모형: k 개 고려하여) :

$$F = \frac{[SSE(RM) - SSE(FM)] / (k - (p+1))}{SSE(FM) / (n - (p+1))} \rightarrow k \leq p+1 \text{ 일 때 } F \text{는 } F\text{-분포이다}$$

H_0 를 기각 $F \geq F_{(p+1), k, n-(p+1), \alpha}$

H_0 를 채택 $F < F_{(p+1)-k, n-(p+1), \alpha}$

3.10.1 회귀계수들이 0인가.

$$\textcircled{1} \quad H_0: \beta_1 = \beta_2 = \dots = \beta_p \quad \text{vs} \quad H_1: \beta_1, \beta_p \text{ 중 } 0 \text{ 이 아닌 } 2 \text{ 가 있다}$$

$$\textcircled{2} \quad H_0(\text{RM}): Y = \beta_0 + \varepsilon \quad \text{vs} \quad H_1(\text{FM}): Y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon$$

이때 RM은 $k=1$ 인 모형이 되며 회귀계수가 있는 모형으로 $\hat{Y}_i^* = \bar{y}$ 가 된다.

$$\text{즉 } SSE(\text{RM}) = \sum (y_i - \bar{y}_i)^2 = SST \quad \text{가지고 } SSE(\text{FM}) = SSE \text{ 이다}$$

$$F = \frac{\frac{SSE(\text{RM}) - SSE(\text{FM})}{p+1-k}}{\frac{SSE(\text{FM})}{n-p-1}} = \frac{\frac{SST - SSE}{p}}{\frac{SSE}{n-p-1}} = \frac{\frac{SSR}{p}}{\frac{SSE}{n-p-1}} = \frac{MSR}{MSE} \quad (\text{평균회귀계수})$$

R_p^2 을 표한 대체상관계수라 할 때 F 결정값은 다음과 같이 표현 가능하며 위의 값과 일치한다.

$$F = \frac{R_p^2/p}{1 - R_p^2/(n-p-1)}$$

Source	Sum of Squares	df	Mean Square	F-test
Regression	SSR	p	$MSR = \frac{SSR}{p}$	$F = \frac{MSR}{MSE}$
Residuals	SSE	n-p-1	$MSE = \frac{SSE}{n-p-1}$	

$$SST(\text{전체변수}) = SSR(\text{예측변수에 의해 설명되는 부분}) + SSE(\text{예측변수에 의해 설명되지 않는 부분})$$

3.10.2 회귀계수들의 부정합 0인가.

가능한 많은 수의 모수들을 사용하여 관찰된 현상을 모사하는 것이다. (모사의 균형성)

가장 먼저 모수 $(x_1 \sim x_6$ 를 봄) 개별 결정에서 x_1, x_3 만 유의하고 경고가 나왔다.

$$\text{이때 } H_0: \beta_2 = \beta_4 = \beta_5 = \beta_6 = 0 \quad \text{vs} \quad H_1: \text{not } H_0$$

$$\text{또는 } H_0(\text{RM}): Y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \varepsilon \quad \text{vs} \quad H_1(\text{FM}): Y = \beta_0 + \beta_1 x_1 + \dots + \beta_6 x_6 + \varepsilon$$

는 결정되는 것이다.

결과.

Source	Sum of square	df	Mean Square	F-test
Regression	3042	2	1521.16	32.1
Residuals	1254	21	46.4485	

$$F = \frac{\frac{[SSE(\text{RM}) - SSE(\text{FM})]}{(p+1-k)}}{\frac{SSE(\text{FM})}{n-(p+1)}} = \frac{(1254 - 1149) / 5}{(1149) / 23} = 0.528$$

$$F_{5, 23, \alpha=0.05} = 2.8 \quad \text{즉} \quad F < F_{5, 23, 0.05} = 0.528 \quad H_0 \text{는 거짓} \quad \text{즉} \quad x_1, x_3 \text{는 영향을 미친다.}$$

여기서 R_p^2 는 FM의 표본 대체상관계수, R_q^2 는 RM의 표본 대체상관계수이다.

$$F = \frac{(R_p^2 - R_q^2) / p-q}{(1-R_p^2) / n-(p+1)} \quad \text{표현 가능}$$

* RM과 FM 보다 단항내의 계수 β_1 를 틀기지는 경우

$$H_0: \beta_1 = 0 \quad vs \quad H_1: \beta_1 \neq 0$$

F검정값 $F = t_1^2$ 을 $F_{1, n-p-1, \alpha} = t_{n-p-1, \alpha}^2$ 와 비교하는 것이다.

즉 개별 회귀계수와 같은 결론을 얻을 수 있다.

* $P=1$ 인 경우 단항회귀가 되므로 RM의 회귀계수의 수는 0이 되고 이때의 통계량은

$$t_1 = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)} \quad \text{가 된다}$$

3.10.3 회귀계수들의 동일성

앞의 예제에서 $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon$ (FM) 또는 결과를 보면, 0(혹은 $\beta_1 = \beta_3 = \beta_1'$)라는 가정을 해보자.

$$Y = \beta_0 + \beta_1'(X_1 + X_3) + \varepsilon \quad (\text{RM}) \quad Y = \beta_0 + \beta_1'W + \varepsilon \quad (W = X_1 + X_3)$$

이를 검정해보면 단항의 계수를 틀기지는 경우라고 볼 수 있다

$$F = \frac{(R_p^2 - R_q^2) / p-q}{(1-R_p^2) / n-(p+1)} = 0.365 < F_{1, 21, 0.05} = 4.21$$

즉

$$t^2 = \left[\frac{\hat{\beta}_1 - \hat{\beta}_3}{\text{se}(\hat{\beta}_1 - \hat{\beta}_3)} \right]^2 < F_{1, 21, 0.05} = t_{21, 0.05}^2$$

을 통해 H_0 를 기각하기 않는다. 즉 X_1 과 X_3 가 비슷함에 증명하는데 볼 수 있다.

3.10.4 제약조건이 있는 경우

FM: $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon$ 0(혹은 $\beta_1 + \beta_3 = 1$)이라는 제약조건을 걸어보자.

$$H_0: \beta_1 + \beta_3 = 1 \quad vs \quad H_1: \beta_1 + \beta_3 \neq 1$$

$$\text{또는 } H_0 \text{ (RM): } Y = \beta_0 + \beta_1 X_1 + (1-\beta_1) X_3 + \varepsilon \quad vs \quad H_1 \text{ (FM): } Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon$$

$$\text{이때 RM: } Y = \beta_0 + \beta_1 X_1 + (1-\beta_1) X_3 + \varepsilon \Rightarrow Y - X_3 = \beta_0 + \beta_1(X_1 - X_3) + \varepsilon \Rightarrow Y' = \beta_0 + \beta_1 Y + \varepsilon$$

이걸 검정해보면 다음과 같다

$$F = \frac{(R_p^2 - R_q^2) / p-q}{(1-R_p^2) / n-(p+1)} = 1.62 < F_{1, 21, 0.05} = 4.21 \xrightarrow{\text{즉}} H_0 \text{ 채택. } \beta_1 + \beta_3 = 1 \text{ 0(혹은 } \beta_1 = \beta_3 = 0.5\text{)을 증명한다}$$

마지막 $\beta_1 = \beta_3 = 0.5$ 인 결론은 진행할 수 도 있다

3.11

예측

예측한계

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \dots + \hat{\beta}_p x_{0p}$$

$$y_0 \in (\hat{y}_0 \pm t_{n-(p+1), \alpha/2} \cdot s.e.(y_0)), \quad s.e.(y_0) = \sqrt{1 + X_0^T (X^T X)^{-1} X_0}$$

신뢰한계

$$\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \dots + \hat{\beta}_p x_{0p}$$

$$\mu_0 \in (\hat{\mu}_0 \pm t_{n-(p+1), \alpha/2} \cdot s.e.(\hat{\mu}_0)), \quad s.e.(\hat{\mu}_0) = \sqrt{X_0^T (X^T X)^{-1} X_0}$$

예측한계가 신뢰한계보다 범위가 넓은 것은 당연하다.

3.12 험결을 이용한 대푯값 표현

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_p \end{bmatrix}, \quad E(\varepsilon) = 0, \quad \text{Var}(\varepsilon) = E(\varepsilon \varepsilon^T) = \Sigma^2 I_n$$

라고 하면 $Y = X\beta + \varepsilon$ 라고 표현이 가능하다.

$$E(Y) = E(X\beta + \varepsilon) = X\beta$$

이제 β 의 회귀계수 추정량 $\hat{\beta}$ 를 찾자

$$S(\beta) = E(\varepsilon\varepsilon^T) = (Y - X\beta)^T (Y - X\beta) \quad \text{이 최소화되도록 한다. 이는 } \beta \text{가 평균이 } 0 \text{인 조건하에}$$

$$(X^T X) \hat{\beta} = X^T Y \quad \text{라고 할 수 있고 이를 증명할 수 있다.}$$

$$\text{만약 } (X^T X) \text{가 역행렬을 가지는 경우 } \hat{\beta} = (X^T X)^{-1} X^T Y \quad \text{라고 표현이 가능하다.}$$

 $\hat{\beta}$ 는 Y 에 대한 선형변수로 끌 수 있다. Y 에 대한 적합 \hat{Y} 는 다음과 같다.

$$\hat{Y} = X\hat{\beta} = X \cdot (X^T X)^{-1} X^T Y = PY, \quad P \text{를 모자행렬 or 사영행렬이라고 한다.}$$

진차 백터는 다음과 같다

$$e = Y - \hat{Y} = Y - PY = (I_n - P)Y$$

회귀제곱총량 (R^2) 가 가지는 성질

$$E(\hat{\beta}) = \beta, \quad \text{Var}(\hat{\beta}) = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] = \Sigma^2 (X^T X)^{-1} = \Sigma^2 C, \quad \hat{\beta} \text{는 } \beta \text{의 Blue 리그이다.}$$

진차제곱률의 특징으로 (역정: $A^T = A$, 역등: $A^2 = A$)

$$e^T e = [I_n - P] Y^T [I_n - P] Y = Y^T (I_n - P) Y \quad \text{즉 } (I_n - P) \text{는 역정이면서 역등의 성질이 있다.}$$

$$\text{방식 } \Sigma^2 \text{의 분산총량은 } \hat{\Sigma}^2 = \frac{e^T e}{n-(p+1)} = \frac{Y^T (I_n - P) Y}{n-(p+1)}$$

$$\hat{\beta} \sim MVN(\hat{\beta}, \Sigma^2 C_{jj}) \Rightarrow \hat{\beta}_j \sim N(\beta_j, \Sigma^2 C_{jj}) \quad \text{라고 할 수 있으므로 } \beta_j \text{의 표준오차는 } \sqrt{\Sigma^2 C_{jj}}$$

$$W = \frac{e^T e}{\sigma^2} \sim \chi^2(n-(p+1)) \quad / \quad \hat{\beta} \text{와 } \sigma^2 \text{은 서로 독립이다}$$

$$\hat{Y} \sim MVN(X\hat{\beta}, \sigma^2 I_p)$$

$$e \sim MVN(0, \sigma^2(I_n - X\hat{\beta}))$$

예측한계 $\hat{y}_0 \in (\hat{y}_0 \pm t_{n-(p+1)} \cdot se(\hat{y}_0))$, $se(\hat{y}_0) = \sqrt{1 + x_0^T (X^T X)^{-1} x_0}$

신뢰한계 $\hat{\mu}_0 \in (\hat{\mu}_0 \pm t_{n-(p+1)} \cdot se(\hat{\mu}_0))$, $se(\hat{\mu}_0) = \sqrt{x_0^T (X^T X)^{-1} x_0}$

회귀모수 β 에 대한 $100(1-\alpha)\%$ 경합 신뢰영역은 다음과 같은 디플로트로 주어진다.

$$\left| \beta: \frac{(\beta - \hat{\beta})^T (X^T X) (\beta - \hat{\beta})}{\sigma^2 (p+1)} \leq F_{p+1, n-(p+1), \alpha} \right|$$

4장 회귀진단

4.1 소개

추론은 자료로부터 구할 수 있는 모양통계량들에 근거하여 진행된다. 그런데 위 가정은 회귀분석의 기본 가정이 만족할 때 그 결과가 유효하다. 즉 기본 가정이 만족하지 않는다면 결과에 심각한 오류를 야기할 수도 있다.

회귀진단은 기본 가정에 위배되거나 치환하는 것이다.

4.2 회귀분석의 표준적인 가정들

1) 모형에 대한 가정

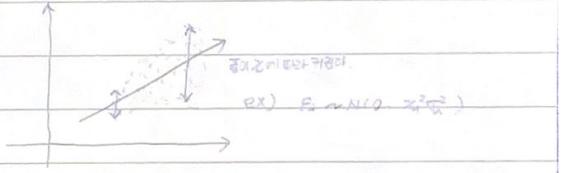
설명변수 X 와 회귀계수 β 를 선형으로 결합하여 반응변수 Y 를 설명할 수 있다고 가정한다.

$$\text{즉, } Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} + \varepsilon_i, \quad i=1, 2, \dots, n \Rightarrow \text{이를 선형성 가정이라 한다.}$$

제크하는 방법으로 신점도를 그려볼 수 있는데, 단순회귀에선 쉽지만 다중회귀 분석에서는 험하다.
선형성이 만족하지 않으면 때때로 자료를 변환시키거나 선형성을 만족시키게 만든다.

2) 오차에 대한 가정

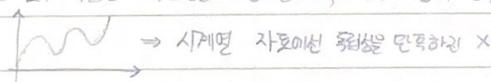
$$\varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$



✓ 정규분포를 따른다는 점에서 정규성을 가정

✗ 분산이 모두 같은 크기의 σ^2 를 갖는다고 하는데 이를 등분산이라고 한다. 만약 분산이 다르면 미분산 문제라고 한다.

✗ 오차들은 모두 독립인 독립성을 가정하는데, 이 가정이 만족하지 않으면 고지상관의 문제가 있다고 부른다.



3) 예측변수에 대한 가정

무작위성이라면 추정성이가 훨씬다.

- 예측변수들은 (X_1, X_2, \dots, X_p) 는 무작위성이 없다. 즉, 미리 고정되어 있는 변수라고 볼 수 있다. 이 가정은 실제로 평가하는 힘들다.

- 예측변수들은 오차가 없이 측정되었다고 가정한다. 이 가정 또한 실제로 평가하는 힘들다.

✓ 예측변수들은 선형 특성이 아니라 선형 특성이 가정한다. 이는 회귀분석이 유용함을 보는데 필요하다.

정규방정식 $(X^T X) \hat{\beta} = X^T Y$ 에서 X 가 선형독립일 때 $(X^T X)$ 가 역행렬을 갖게 되면서 $\hat{\beta} = (X^T X)^{-1} X^T Y$ 를 계산된다.

이 가정에 위배가 되면 공정성이 문제가 있다고 말한다. 예) Y : 가임질수 X_1 : 2주기 고정 $I_m(Y \sim (X_1 + X_2)) \Rightarrow \text{OK}$
 X_1 : 고체질수 X_2 : 고체질 $I_m(Y \sim (X_1 + X_2)) \Rightarrow \text{오류}$

공정성이 어려워지거나 된다.

4) 관측개체에 대한 가정

모든 관측 개체들은 동일하게 선형하는 만 하며, 회귀 결과를 결정하고 결론을 도출함에 있어 거의 등등한 역할을 한다.

4.3 다양한 유형의 진차들

가장 기본적인 방법은 진차그림을 살펴보는 것이다.

선형모형에서 회소제곱법에 의해 적합값을 얻는다면 다음과 같다.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_p x_{ip} \quad i=1, 2, \dots, n$$

보통의 회소제곱진차는 다음과 같다. $\Rightarrow e_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_p x_{ip})$ (구현식 X)

$$e_i = Y_i - \hat{Y}_i, \quad (i=1, 2, \dots, n) \quad e_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_p x_{ip}) \quad (\text{구현식 O})$$

적합값은 다음과 같이 다시 쓸 수 있다.

$$\hat{Y}_i = P_{ii} Y_1 + P_{i2} Y_2 + \dots + P_{in} Y_n, \quad i=1, 2, 3, \dots, n \quad (P_{ij} \text{은 } 0 \text{이 아닌 변수를 위한 관계 있는 양으로 반복변수와 같은 } P_{ii} \text{는 } 1 \text{이다.})$$

단순회귀모형에서 P_{ij} 는 다음과 같다.

$$P_{ij} = \frac{1}{n} + \frac{(x_{i1} - \bar{x})(x_j - \bar{x})}{\sum_k (x_k - \bar{x})^2}, \quad \text{이후 다른 선형회귀방식에서는 모자행렬 or 사슬행렬 } P = X(X^T X)^{-1} X^T \text{와 같은 것을 사용한다.}$$

$$X = \begin{bmatrix} 1 & x_{i1} & x_{i2} & \dots & x_{ip} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}, \quad \hat{Y} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix} = X \hat{\beta} = [X(X^T X)^{-1} X^T] Y - PY$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad , \quad = \begin{bmatrix} P_{11} & P_{12} & \dots & P_{1n} \\ \vdots & \vdots & & \vdots \\ P_{n1} & P_{n2} & \dots & P_{nn} \end{bmatrix} \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix} \Rightarrow \hat{Y} = P_{11} Y_1 + \dots + P_{nn} Y_n$$

특히 $i=j$ 일때 P_{ii} 는 P 의 i 번째 대각원소가 되면서 $P_{ii} = \frac{(x_{ii} - \bar{x})^2}{\sum_k (x_k - \bar{x})^2}$ 이된다. 이것은 i 번째에 대한

지지값 (가장값과 비슷한 것들)이라고 한다. 따라서 n 개의 지지값이 존재함에 따라 $P_{11}, P_{22}, \dots, P_{nn}$ 으로 층의되어 귀중한 역할을 한다.

앞서 가정을 만족하면 진차 ($e_1 \sim e_n$)의 합은 0이며 불산을 따르면 $e_i = Y - \hat{Y} = Y - PY = (I_n - P)Y = \Xi$

$$\text{Var}(e_i) = \sigma^2 (1 - P_{ii}) \text{로 } x_{i1}, x_{i2}, \dots, x_{ip} \text{에 의존성이 떨어진다.} \quad e_i \sim N(0, \sigma^2 (1 - P_{ii}))$$

여기서 P_{ii} 를 표준화 하게 되면 i 번째 표준화진차라고 하며 $Z_i \sim N(0, 1)$ 이 된다.

$$Z_i = \frac{e_i - 0}{\sqrt{\sigma^2 (1 - P_{ii})}} = \frac{e_i}{\sigma \sqrt{(1 - P_{ii})}}, \quad \text{여기서 } \sigma \text{는 모르겠지만 표본을 통해 추정해보자.}$$

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-(p+1)} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-(p+1)} = \frac{\text{SSE}}{n-(p+1)}, \quad E(\hat{\sigma}^2) = \sigma^2$$

이번에는 오차 불린 $\hat{T}_{(1)}^2$ 에 대해서 다른 측정값을 보자. n 개의 자료가 아니라 i 번째 자료가 없는 ($n-1$)개의 자료에서

$$\hat{T}_{(1)}^2 = \frac{SSE_{(1)}}{(n-1)-(p+1)} = \frac{SSE_{(1)}}{n-p-2}, \quad E(\hat{T}_{(1)}^2) = T^2 \text{ 이성립한다.}$$

$SSE_{(1)}$ 는 i 번째를 제외한 나머지 진짜 평균합이며, $Var(\hat{T}^2) < Var(\hat{T}_{(1)}^2)$ 가성립한다.

측정변수가 더 적어서

- 내적 표준화 진짜: $r_i = \frac{e_i}{\sqrt{\sum_j 1 - p_{ij}}}$
- 외적 표준화 진짜: $r_i^* = \frac{e_i}{\sqrt{\sum_{j \neq i} 1 - p_{ij}}}$

\Rightarrow 자료에 이상점이 있으면 측정값이 끌어쓰게 나오지 않을 수 있다. 따라서 이상치를 뺀 고산한 $\hat{T}_{(1)}^2$ 을 사용하는 게 옳다.

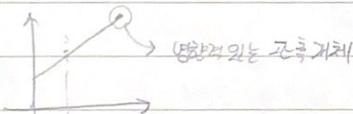
표준화된 진짜들은 합계가 0이 아니지만 평균은 1이다. 외적 표준화진짜는 $T(n-p-2)$ 를 고르지만 내적 표준화진짜는 그렇지 않는다. 하지만 표본을 엄정화하면 두 진짜들은 표준정규 분포를 따르게 되며, 진짜간 독립이 아니었던 독립성의 문제도 해결 되었다.

두 r_i^* 와 r_i 의 관계는 $r_i^* = r_i \sqrt{\frac{(n-1)-(p+1)}{n-(p+1)-\hat{T}_{(1)}^2}}$ 다음과 같고 우리는 주로 내적 사용을 한다.

4.4 그래프적 방법들

Anscombe's Quartet Data.

4개의 자료의 $Cor(X, Y)$ 는 거의 비슷하지만 양상을 다름. 즉 상관예수와 그래프를 같이 구하고 결론을 내리는 게 좋다.



그래프적 방법은 여러방식으로 이용될수 있다. (자료에 따른 경과, 특이값 찾기, 더 나은 분석을 위한 방법, 새로운 현상 찾기 등)



4.5 모형을 접합함에 이전의 그래프

X와 Y의 관계를 나타내는 모형 접합에 대해 그래프를 살펴보는건 텁색적 도구의 역할을 한다

1) 일차원 그래프

히스토그램, 즐기와 막그림, 점体质, 네트워크

1. \Rightarrow 분포를 알 수 있으며, 치우침이 심하면 로그변환이 추천된다.

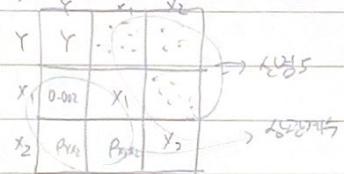
2. \Rightarrow 특이값의 존재 유무를 알 수 있다. 특이값이 존재하면 분석시 주의깊게 다뤄야 한다.

2) 이차원 그리프

다차원 자료가 주어질 때, 변수가 적다면 A를 쌍별로 신경도를 행렬로 넣어둘 수 있다.

레도사의 풀트 or 선형도형렬 (Psych pair panels (data))?

선형관계만 허용 가능하여, 로버스트 (한 두 관측자체의 영향을 고려 받지 않는다) 하지 않는 이유로 신경도와 같이 허용되며 주. 상관계수가 크다고 꼭 선형은 아니기 때문에 그림과 같이 비교해보는 게 좋다.



4.6 모형을 접합하는 이후의 그래프

모형접합 이후의 그래프는 가정을 검토하고 주어진 모형의 적합도를 평가하는데 도움을 준다.

- 선형성과 정규성 가정을 검토하기 위한 그래프
- 특이값과 영향력 있는 개체를 검출하기 위한 그래프
- 변수들의 효과에 대한 진단들을

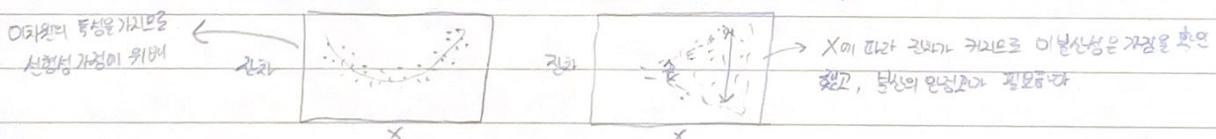
4.7 선형성과 정규성 가정에 대한 검토.

- 표준화 진차의 정규성을 그림 : 순서화된 표준화진차 ($\gamma_{i(1)} = \frac{e_i}{\sqrt{1-p_{ii}}}$) 데 정규집수 (표준정구 분포에서 $(Q-Q plot)$ \rightarrow e_i 가 n 인 표분을 최적분을 때 일어날 것으로 추정되는 값) 의 풀트이다.

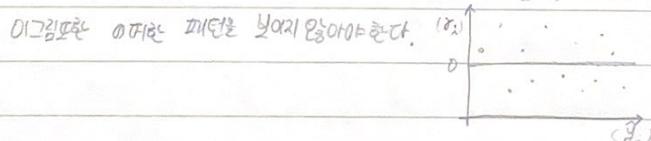
정규성을 만족하면 $y = x$ 의 그래프를 가져야 한다.

- 표준화 진차에 각 예측 변수들의 신경도 : 표준적으로 표준화진차는 각 예측 변수들과 상관되어 있지 않는데 이 가정이 만족되면

진차 vs 예측변수 그림에는 아득한 패턴이 없어야 한다.



- 표준화진차 대 적합값의 풀트 : 표준화된 가정하에서 표준화진차는 적합값과 상관되어 있지 않다. 따라서 이 가정이 만족된다면



- 표준화진차의 인덱스 풀트 : 인덱스 풀트은 표준화진차에 고속 기체 빙고의 풀트를 의미한다. 시계열자료와 같이 순서가 중요한 자료는 연속적인 순서를 고려하여 진차 그림을 그리고, 오각의 득점설 검정을 검토해야 한다.