Chapter 2 단순회귀

회귀 분석은 종속변수(Y)와 설명변수들($X_1, X_2, ..., X_p$, 독립변수)과 관계를 분석하는 도구이다. (1)모형에 설정된 설명변수들의 유의성 검정?(모형과 회귀계수의 유의성 검정) (2) 유의한 설명변수 중 종속변수에 영향력이 가장 큰 변수는 무엇인가?(표준화 회귀계수) (3) 그리고 설명변수 값들이 주어진 경우 종속변수의 예측치는? 이에 대한 해답을 회귀분석이 제공한다.

회귀 분석은 종속변수(Y)와 설명변수들($X_1, X_2, ..., X_p$, 독립변수) 사이의 함수는 매우 다양하다. $Y = f(X_1, X_2, .., X_p)$ 일반적으로 다루기 편하고 해석이 용이한 선형함수 형태를 고려하게 되는데 이런 이유로 회귀분석은 선형회귀분석이라고도 한다. 선형함수가 아닌 회귀모형에 대한 분석을 비선형(nonlinear) 회귀분석이라 한다. 비선형회귀모형에 대한 분석은 다소 복잡하고 무엇보다도 모형에 대한 해석이(선형회귀모형에서 회귀계수의 의미는 편미분 계수이므로 해석이 용이) 쉽지 않아 사용 빈도가 낮다.

선형회귀 모형: $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{p1i} + e_i$, $i = 1,2,...,n$, $e_i \sim iidN(0, \sigma^2)$

선형변환 경제학의 Cobb-Douglas 생산 모형은 비선형 모형이다. $Q = \alpha K^{\beta} L^{\lambda} u$, Q=생산량, K=자본, L=노동력, u는 오차항이고 나머지는 계수이다. 양변에 Log를 취하면 $\ln(Q) = \ln\alpha + \beta \ln K + \lambda \ln L + \ln u$ 이 되므로 선형 회귀 분석이 가능하다.

인구성장모형 $P_t = \alpha e^{\beta T} e_t$도 선형변환이 가능하다. $\ln Q_t = \ln\alpha + \ln(\beta) \times T + \ln(e_t)$

회귀계수 선형 회귀 모형에서 회귀계수(β_k)의 의미는 설명변수 X_k가 한 단위 증가할 때 종속변수가 얼마나 변하는지(편미분 계수) 나타내는 값이다. 이처럼 선형 모형에서는 회귀계수에 대한 해석이 용이하다.

회귀모형에서 설명변수가 하나인 경우에 대한 분석을 단순회귀분석이라 하는데 이 장에서는 회귀분석에 대한 개념, 이론, 분석 방법에 대한 이해를 높이기 위하여 이를 살펴보고자 한다.

2.1 산점도

2.1.1 개요

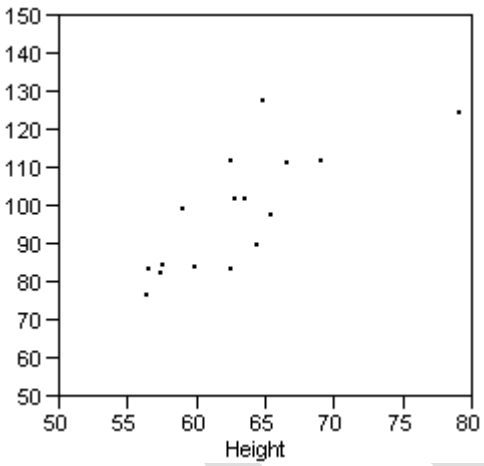
산점도(scatter plot)는 두 변수간의 (함수) 관계를 나타내는 이차원 그래프로 종속변수는

Y축, 설명변수는 X축으로 (인과 관계가 존재할 때, 상관 관계 존재할 때는 아무 변수나 Y축에 지정) 하여 데이터 관측치 쌍을 그린다. 산점도를 통해 두 변수 (X, Y) 간의 함수 관계를 쉽게 파악할 수 있으므로 회귀분석의 시작이다.

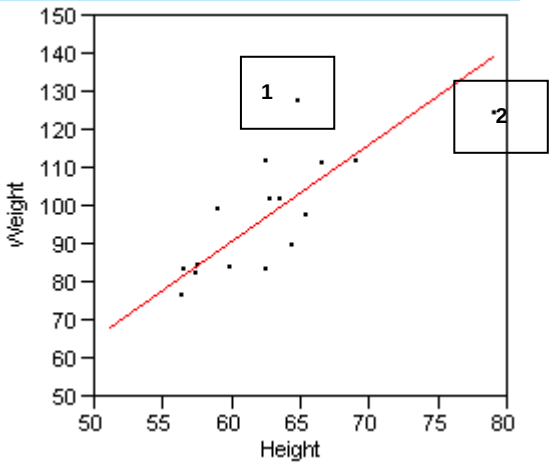
다음은 키와 몸무게(종속변수)의 인과 관계를 알아보기 위해 조사한 자료의 일부이다.

	Name	Height	Weight
1	BARBARA	65.3	98
2	ALFRED	69	112.5
3	ALICE	56.5	84

표본 18명의 키와 몸무게 데이터에 대한 산점도를 그린 것이다. (그림1) 만약 산점도를 그리지 않고 F-검정에 의한 분산 분석만(산점도 아래)을 실시하면 설정한 회귀 모형은 적절하면 이로부터 추정된 모형을 이용하게 된다. Weight = -63.626 + 2.56828 Height



[그림 1]



[그림 2]

그러나 산점도를 살펴 되면(그림2, 물론 붉은 선은 추정 회귀선) 두 가지 특이한 관측치가 발견된다. 1 관측치는 같은 키의 다른 사람에 비해 몸무게가 많이 나가는 것을 알 수 있다.[이상치 outlier] 2 관측치는 키와 몸무게의 관계가 선형이 아니라 2차식 관계가 성립하지 않을까 하는 의심을 갖게 한다.[영향치 influential observation] 이처럼 산점도는 다음을 알 수 있다.

①종속변수와 설명변수들간의 함수(직선) 관계(유의성)를 미리 진단할 수 있다.

②설명변수들간의 상관 관계가 존재하는지 알 수 있어 다중공선성 문제를 예상할 수 있다.다중공선성 문제는 다중회귀모형에서 일어나는 일이다. 다중회귀분석에서는 산점도 행렬(scatter plot matrix)을 그리게 된다.

③특이한 관측치(이상치, 영향치)가 존재하는지를 알 수 있다.

2.1.2 산점도 그리기

EXAMPLE 2-1 산점도 그리기

AD.xls (엑셀 데이터)

1983년 미국 21개 기업 광고비(SPEND, 단위: 백만$)가 소비자 평가도(RATE)를 조사한 것이다. 두 변수간의 함수 관계를 살펴보기 위하여 산점도를 그려보자.

광고비와 소비자 평가간의 함수 관계가 있는지 알아보기 위하며 산점도를 그려보자.

AD.xls

	A	B	C	D
1	FIRM	SPEND	RATE	GROUP
2	MILLER LITE	50.1	32.1	1
3	PEPSI	74.1	99.6	1
4	STROH'S	19.3	11.7	2

데이터 입력이 정형화 되지 않은 경우 INFILE 문은 사용해 SAS 데이터를 만드는 것은 시간 낭비이다. 회귀분석 데이터의 경우 데이터의 양이 많지 않으므로 CTRL+C▶CTRL+V를 이용해 프로그램 에디터(확장 편집기)에 데이터를 복사하여 DATALINES;(혹은 CARDS;) 이용하여 SAS 데이터를 만들자. 1-14의 의미는 1열부터 14열까지를 변수 name의 관측치로 읽으라는 명령이다. 만약 이 옵션을 사용하지 않으면 8자리만 읽어온다.

SAS

```
data ad;
    input name $ 1-12 spend rate group;
    datalines;
PEPSI        74.10   99.60   1.00
STROH'S      19.30   11.70   2.00
FED'L EX     22.90   21.90   2.00
```

▶

```
proc gplot data=ad;
    symbol i=rl v=circle;
    plot rate*spend;
run;
```

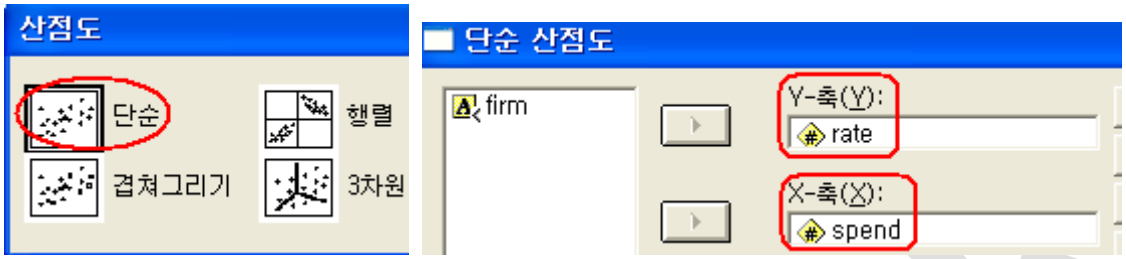
```
proc reg data=ad;
    model rate=spend;
    plot rate*spend;
run;
```

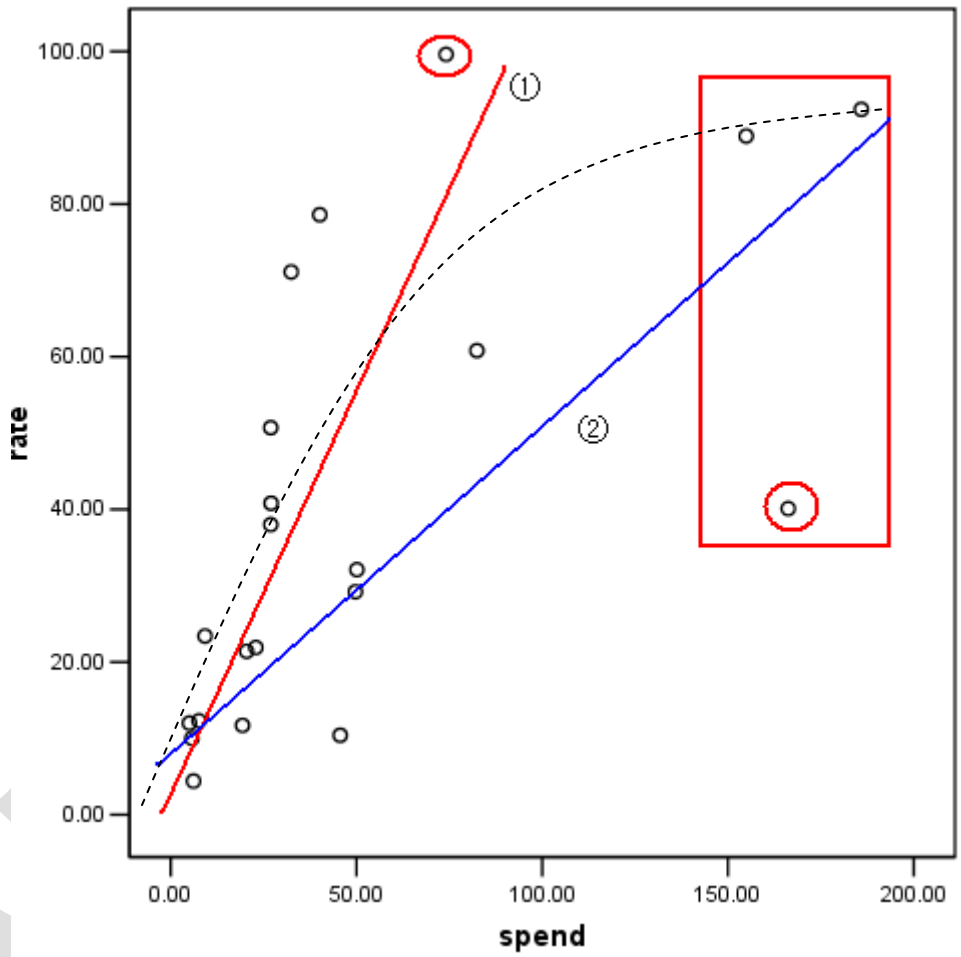
(이전 프로그램)

▷I=Interpolation 관측치를 연결하는 방법이다. JOIN은 관측치를 직선으로 연결하는 것이고 SPLINE는 곡선으로 연결한다. R은 회귀선(Regression Line)을 의미하며 L은 직선(Linear)을 의미한다. 이차 곡선이면 L 대신 Q를 적어주면 된다.

▷V=Value 관측치 점의 모양 설정한다. dot, triangle, square, star 등을 사용할 수 있고 V='a' 사용하면 점들이 a로 찍힌다.

SPSS 메뉴에서 그래프(G) ▶ 산점도(S)... 선택하고 다음과 같이 메뉴를 설정한다.





광고비가 증가할수록 평가 정도는 높아짐을 알 수 있다. 원의 관측치는 이상치(outlier)일 가능성이 높다. 문제는 네모 상자 안에 있는 관측치들이다. 이 관측치들로 인하여 특히 원 안의 관측치, 추정된 회귀직선은 ②와 같다. 그러나 네모 상자 안의 관측치 3개를 제외하면 추정회귀식은 ①이다. 어느 것이 적절한가? 광고비를 많이 지출하는 회사를 보니 자동차 회사 FORD, 전화 회사 ATT, 그리고 MacDonald이다. 이들 회사는 다양한 이유로 유난히 많이 광고비를 지출하고 있는 회사이다. 그러므로 이들을 제외하고(분석자 판단) 회귀모형을 추정하는 것이 바람직해 보인다.

만약 네모 안의 3 관측치가(그렇다면 원의 관측치는 이상치이지만) 유효하다면 직선 관

계가 어니라 이차식을 관계가 아닐는지? 즉 점선의 이차식으로 광고비용이 평가도에 영향을 미치고 있다고 해야 하는 한다.

HOMEWORK #2-2 DUE 3월 16일(수)

CARPRICE.txt (텍스트 데이터)

1990년 미국 Ford 자동차 구매자 중 62명을 임의 선택 조사한 자료이다. 목적은 구매하는 자동차 종류(가격)에 영향을 미치는 변수(요인)를 알아보기 위한 것이었다. 성별(1=남자, 0=여자), 연 소득($), 나이, 결혼여부(1=기혼, 0=미혼), 자녀 수, 학력(1=대졸, 0=고졸이하), 자동차 가격($) 구매하는 자동차 가격에 연 소득이 영향을 미치는지 알아보고자 회귀분석을 실시하고자 한다. 먼저 산점도를 그리고 해석하시오. (SAS 이용하기)

2.2 모형 및 가정

2.2.1 모형

종속변수를 y, 설명변수를 x라 하고 첨자(subscript) i는 관측치를 나타내며 n을 표본의 개수라 하면 선형 회귀 모형(model)은 다음과 같다.

$$Y_i = \alpha + \beta X_i + e_i, \quad i = 1,2,...,n \quad ---(1)$$

- 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 계수이다.
- x: 설명변수, 확률변수가 아닌다.
- e:오차항(error term), 회귀직선($\alpha + \beta x_i$)에 의해 설명되지 못하는 부분

모형 (1)을 관측치에 따라 풀어 쓰면 다음과 같다.

$$\begin{aligned} y_1 &= \alpha + \beta x_1 + e_1 \\ y_2 &= \alpha + \beta x_2 + e_2 \\ &... \\ y_n &= \alpha + \beta x_n + e_n \end{aligned}$$

모형 (1) 행렬의 형태로 표시하면 $\underline{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = X\underline{\beta} + \underline{e}$

2.2.2 가정

회귀 모형의 가정(assumption)은 다음과 같다.

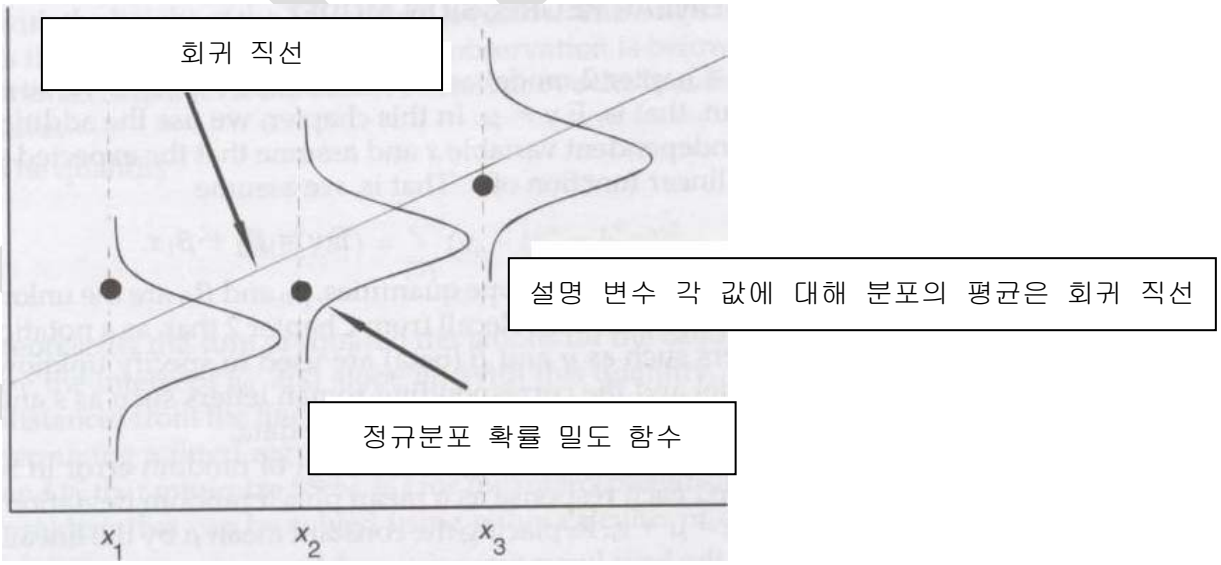
①회귀계수 α, β 는 모수이며 상수(constant)이다.

②종속변수와 설명변수 간에는 선형(직선) 함수 관계가 존재한다.

③설명변수 X 는 확률변수가 아니라 수학변수(deterministic)로 오차 없이 측정할 수 있다고 가정한다. 그러므로 회귀모형에서 확률변수는 e 와 Y 이다. 확률변수이면 확률분포함수를 갖는다.

④ $e_i \sim iid\ N(0, \sigma^2)$: independently and identically distributed

- **독립성(independent)**: 오차항은 서로 독립이다. 즉 각 오차는 서로 영향을 주지 않는다. 독립성 가정은 시계열 데이터(시간적 순서를 갖는 데이터) 경우에만 체크한다.
- **정규성(normality)**: 오차항은 정규 분포를 따른다. 이 가정은 F-검정 방법을 사용하기 위하여 반드시 필요하다.
- **등분산성(homoscadicity)**: 오차항의 분산은 동일하다. 분산이 일정하다는 가정의 주어진 설명변수 값에서 관측되는 Y 의 값의 분산이 일정하다는 의미와 같다. 분산이 다르면 설정된 회귀 모형이 적절함에도 불구하고 관측치가 직선에 모여 있지 않게 된다. 분산이 크므로 벗어나는 경향이 있다.



EXAMPLE 2-2 종속변수의 평균과 분산

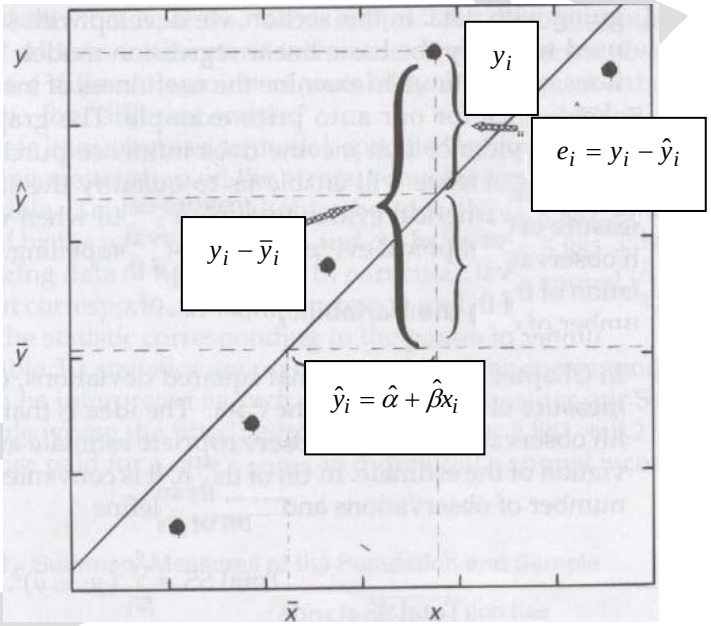
가정 $e_i \sim iid\ N(0, \sigma^2)$ 하에서 종속변수 Y 의 분포, 평균과 분산을 구하시오.

2.3 회귀계수 추정

회귀 모형을 추정한다는 것은 수집된 데이터(산점도)에 가장 적절한 회귀 직선을 구하는 것이다. 방법으로는 OLS(Ordinary Least Square: 최소자승법)과 MLE(MLE: Maximum Likelihood Estimator: 최대 우도 추정법, 최우 추정법) 방법이 있다.

2.3.1 최소자승법

각 관측치에 가장 적합한 회귀 직선은 회귀 직선과 관측치의 벗어난 정도(오차: e_i)가 가장 적은 직선일 것이다. 그런데 $\sum_{i=1}^{n} e_i = 0$ 이므로 $\sum_{i=1}^{n} e_i^2$ (절대값 대신 제곱하는 이유는 (1)다루기 쉽고 (2)멀리 떨어질수록 더 큰 페널티를 부여)을 최소화 하는 α, β 을 추정하는 방법을 최소자승법(OLS)라 한다.



$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$ 을 최소화 하는 추정치 $\hat{\alpha}, \hat{\beta}$ 를 OLS 추정치라 한다. 즉, OLS 추정치를 구하려면 Q를 α, β에 대해 각각 편미분(partial derivative) 하고 그 결과를 0이라 놓고 얻은 연립 방정식을 풀면 된다. 이를 정규방정식이라 한다.

정규 방정식(normal equation):

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

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

정규 방정식에서 α, β의 해를 구하면 다음과 같고 이를 OLS 추정치라 한다.

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

$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}), S_{xx} = \sum (x_i - \bar{x})^2$ 라 정의하면 $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$ 이다.

2.3.2 최대우도 추정량(MLE)

최대우도 추정량이란 우도함수(Likelihood function, 확률 밀도 함수의 곱)를 최대화 하는 추정량이다. 확률표본 $(x_1, x_2, ..., x_n) \sim f(x;\theta)$ 인 경우 우도 함수는 $L(\theta; x_1, x_2, ..., x_n) = \prod f(x_i;\theta)$ 이고 이 함수를 최대화 하는 $\hat{\theta}$을 최대 우도 추정량(MLE)이라 한다.

회귀모형의 가정으로부터 $(y_1, y_2, ..., y_n) \sim N(\alpha + \beta x_i, \sigma^2)$ 임을 알았다. 그러므로 우도함수는

$$\begin{aligned} L(\alpha, \beta; x_1, x_2, ..., x_n) &= f(y_1, y_2, ..., y_n; \alpha, \beta) \\ &= \Pi \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(y_i - \alpha - \beta x_i)^2}{2\sigma^2}) \\ &= (\frac{1}{\sqrt{2\pi}\sigma})^n \exp(-\frac{\sum (y_i - \alpha - \beta x_i)^2}{2\sigma^2}) \end{aligned}$$

우도함수를 최대화 하는 α, β 을 구하면 이것이 MLE 추정치이다. 우도함수를 최대화 한다는 것은 우도함수의 로그 $\ln L \propto \sum (y_i - \alpha - \beta x_i)^2$ 을 최대화 하는 것과 동일하다. 그러므로 회귀계수에 대한 MLE 추정치는 OLS 추정치와 동일하다.

3.3.3 추정된 회귀식 성질

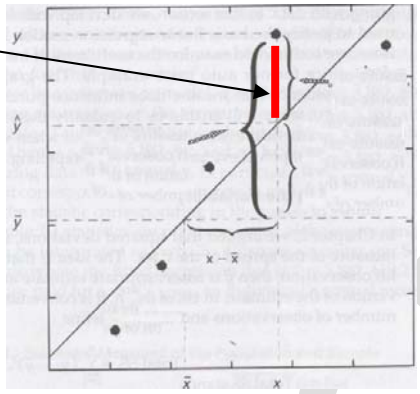
최소자승(OLS) 추정치 $\hat{\beta} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$, $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ 을 이용하여 얻은 $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$ 을 추정된 회귀식 혹은 적합된(fitted line) 회귀직선이라 한다.

GAUSS-MARKOV Theorem
회귀계수에 대한 OLS 추정치는 BLUE(Best Linear Unbiased Estimator)이다. 즉 모든 선형, 불편 추정량 중 최소 분산(minimum variance)를 갖는다.

3.3.2절에 의해 본 것처럼 MLE와 OLS 추정치와 동일하다. MLE 추정치의 성질에 의하면 MLE 함수 중 불편성을 갖는 추정치는 Rao-Blackwell 정리에 의해 그 추정량이 **MVUE** 이다. 다음에 살펴 보겠지만 OLS 추정치는 불편성을 갖는다. 그러므로 GAUSS-Markov 정리가 증명된다.

종속변수의 예측치(적합치) $\hat{y}_i$와 실제 관측치의 차이를 잔차(residual)라 하면 이는 오차의 추정치가 되는데 이는 다음 장에서 다루기로 한다.

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



적합된 회귀직선의 성질을 살펴 보면 다음과 같다.

① $\sum r_i = 0$ 잔차의 합은 0이다.

② $\sum x_i r_i = 0$ (관측치 x_i을 가중치로 한 잔차의 가중치 평균은 0이다)

③ $\sum \hat{y}_i r_i = 0$ (예측치 $\hat{y}_i$을 가중치로 한 잔차의 가중치 평균은 0이다)

④적합된 회귀직선은 $(\bar{x}, \bar{y})$을 지난다.

EXAMPLE 2-3 잔차 성질 증명하기

---

①, ②는 정규방정식에 의해 당연하다.

③ $\sum \hat{y}_i r_i = \sum (\hat{\alpha} + \hat{\beta}x_i) r_i = \sum \hat{\alpha} r_i + \sum \hat{\beta} x_i r_i = 0$ from ①, ②

④ $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$

2.4 회귀계수에 대한 가설검정

단순회귀모형 $y_i = \alpha + \beta x_i + e_i$에서 의하면 α, β는 모수(parameter)이다. 모수에 대한 추정량을 구하는 방법으로 OLS, MLE 방법을 살펴보았고 이는 동일함을 알았다. 이 절에서는 각 추정량의 성질과 분포 함수를 도출하고 모수에 대한 가설 검정하는 방법을 살펴 보기로 하자.

2.4.1 회귀계수 β에 대한 추론

회귀계수 β의 분포

회귀 분석에서 가장 관심을 갖는 것은 기울기 회귀계수 β이다. 설명변수의 유의성을 검

정한다는 것을 귀무가설 $H_0 : \beta = 0$ 의 유의성 검정과 동일하다. $H_0 : \beta = 0$ 채택되면 회귀모형에서 βx_i 가 없어지므로 1)설명 변수(x)는 종속 변수(y)를 설명하지 못하고(유의하지 않다) 2)회귀 모형은 $y_i = \alpha + e_i$ 로 줄어들어 종속변수는 $\hat{\alpha} = \bar{Y}$ 에 의해 설명된다.

β 의 OLS 추정치는 $\hat{\beta} = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$ 이다.

EXAMPLE 2-4 잔차 성질 증명하기

① $\hat{\beta} = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$ 가 $\hat{\beta} = \sum k_i Y_i, k_i = \dfrac{(X_i - \bar{X})}{\sum (X_i - \bar{X})^2}$ 임을 보이시오.

②①의 결과($\hat{\beta}$ 의 y_i 의 선형 결합이다)을 이용하여 $\hat{\beta}$ 의 분포가 정규 분포임을 보이시오.

HOMEWORK #2-3 DUE 3월 16일(수)

(1) $\sum k_i = 0$, $\sum k_i X_i = 1$, $\sum k_i^2 = \dfrac{1}{\sum (X_i - \bar{X})^2}$ 임을 보이시오.

(2)(1)을 이용하여 $E(\hat{\beta}) = \beta$ (불편 추정량), $V(\hat{\beta}) = \sigma^2(\hat{\beta}) = \dfrac{\sigma^2}{\sum (X_i - \bar{X})^2}$ 임을 보이시오.

(3)위의 결과를 이용하여 $\hat{\beta} \sim Normal(\beta, \dfrac{\sigma^2}{\sum (x_i - \bar{x})^2})$ 보이시오.

오차항(e_i)이 정규분포라는 가정(그러면 종속변수) 하에 $\hat{\beta} \sim Normal(\beta, \dfrac{\sigma^2}{\sum (x_i - \bar{x})^2})$ 가 성립한다. OLS 추정치를 구할 때는 오차항의 정규성 가정이 사용되지 않았으나 회귀계수 β 에 대한 가설검정에 사용된다. 대표본 이론이 성립하나? $\dfrac{\hat{\theta} - \theta}{s_{\hat{\theta}}} \underset{app}{\sim} Normal(0,1)$ (no, why?)

Gauss Markov Theorem

회귀계수에 대한 OLS 추정치는 BLUE(Best Linear Unbiased Estimator)이다. 즉 모든 선형, 불편 추정량 중 최소 분산(minimum variance)를 갖는다.

[증명]1 $\hat{\beta} = \dfrac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$ 가 Y_i 선형 추정량이며(Linear) 불편(Unbiased) 추정량 $E(\hat{\beta}) = \beta$

임을 증명하였다. 이제 선형 추정량 중 최소 분산을 가짐을 증명하면 된다. β 의 선형 추정량을 $\hat{\beta}_* = \sum c_i Y_i$ 라 하면 불편 추정량이 되기 위해서는 $\sum c_i = 0$, $\sum c_i X_i = 1$ 을 만족해야 한다. 추정량 $\hat{\beta}_* = \sum c_i Y_i$ 의 분산은 $\sigma^2(\hat{\beta}_*) = \sigma^2 \sum c_i^2$ 이다. 만약 $c_i = k_i + d_i$ 이라 하면 (OLS 추정량의 계수 $k_i = \frac{(X_i - \bar{X})}{\sum (X_i - \bar{X})^2}$) $\sigma^2(\hat{\beta}_*) = \sigma^2(\hat{\beta}) + \sigma^2 \sum d_i^2$ (왜냐하면 $\sum k_i d_i = 0$)이므로 $\sigma^2(\hat{\beta}_*)$ 의 분산은 $\sum d_i^2$ 가 0인 경우이므로 모든 d_i 가 0이어야 한다. 그러므로 OLS는 최소 분산을 갖는 선형 불편 추정량이다.

[증명2] Rao-Blackwell 정리에 의해 MLE의 함수 중 불편성을 갖는 추정치는 MVUE이다. 앞에서 회귀모형의 회귀계수 추정치 OLS는 MLE 추정치와 같고 불편 추정량임을 보였다. *Q.E.D*

$\frac{\hat{\beta} - \beta}{s(\hat{\beta})}$ 의 분포함수(sampling distribution)

앞에서 $\hat{\beta} \sim Normal(\beta, \frac{\sigma^2}{\sum (x_i - \bar{x})^2})$ 임을 알았으나 오차분산 σ^2 을 모르므로 추정해야 한다. 오차에 대한 MVUE 추정량은 $\hat{\sigma}^2 = \frac{\sum (y_i - \hat{\alpha} - \hat{\beta} x_i)^2}{n-2} = MSE$ 이다. 그리고 $\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$ 이므로 $\frac{\hat{\beta} - \beta}{s(\hat{\beta})} \sim t(n-2)$, where $\hat{\sigma}^2(\hat{\beta}) = \frac{MSE}{\sum (X_i - \bar{X})^2}$ 임을 알 수 있다.

EXAMPLE 2-5 잔차 성질 증명하기

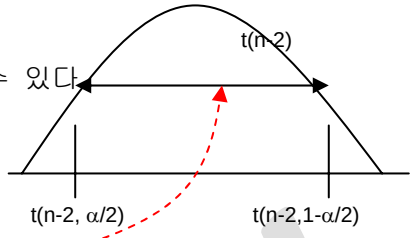
① $\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$ 임을 보이시오. (briefly)

HOMEWORK #2-4 DUE 3월 16일(수)

EXAMPLE2-4 이용하여 $\frac{\hat{\beta} - \beta}{s(\hat{\beta})} \sim t(n-2)$ 임을 보이시오.

β 에 대한 신뢰구간과 가설 검정

$\frac{\hat{\beta}-\beta}{s(\hat{\beta})} \sim t(n-2)$ 을 이용하여 신뢰구간과 가설 검정을 할 수 있다.



①100(1-α)% 신뢰 구간(confidence interval)

$$[\hat{\beta}-t(n-2;1-\alpha/2)*s(\hat{\beta}),\hat{\beta}+t(n-2;1-\alpha/2)*s(\hat{\beta})],\quad s^2(\hat{\beta})=\frac{MSE}{\sum(X_i-\bar{X})^2}$$

신뢰구간에 0이 포함되어 있으면 양측 검정의 경우 $H_0:\beta=\beta_0$ 이 채택된다.

②가설 검정 $H_0:\beta=\beta_0$ (가장 일반적인 형태는 $H_0:\beta=0$ 이다)

검정 통계량 $T=\frac{\hat{\beta}-\beta_0}{s(\hat{\beta})}$ 이다. 만약 양측 검정인 경우($H_0:\beta\neq\beta_0$) 검정 통계량의 절대값이 $t(n-2;1-\alpha/2)$ 보다 크면 귀무가설을 기각하고 단측 검정($H_0:\beta>\beta_0$ 혹은 $H_0:\beta<\beta_0$)인 경우는 $t(n-2;1-\alpha)$ 보다 크면 귀무가설을 기각한다. $\beta_0=0$ (설명 변수가 유의하지 않다)에 관심이 있으므로 양측 검정이 일반적이다.

설명변수가 하나인 단순회귀분석의 경우

2.4.2 회귀계수 α 에 대한 추론

단순 선형 모형에서 α 는 절편에 해당되므로 $H_0:\alpha=0$ 가 채택되면 원점을 지나는 회귀선이 된다. 그러나 일반적으로 절편에 대해 관심이 없으므로 필요한 경우 이외에는 검정하지는 않는다.

$\frac{\hat{\alpha}-\alpha}{s(\hat{\alpha})}$ 의 분포 함수

$\hat{\alpha}=\bar{y}-\beta\bar{x}$ 이므로 $\frac{\hat{\alpha}-\alpha}{\sigma(\hat{\alpha})}\sim Normal(0,1)$, $\sigma^2(\hat{\alpha})=\sigma^2(\frac{1}{n}+\frac{(\bar{X})^2}{\sum(X_i-\bar{X})^2})$ 이다.

$\hat{\sigma^2}(\hat{\alpha})=s^2(\hat{\alpha})=MSE(\frac{1}{n}+\frac{(\bar{X})^2}{\sum(X_i-\bar{X})^2})$ 라 하면 $\frac{\hat{\alpha}-\alpha}{s(\hat{\alpha})}\sim t(n-2)$

α 에 대한 신뢰구간과 가설 검정

①100(1-α)% 신뢰 구간(confidence interval)

$$[\hat{\alpha}-t(n-2;1-\alpha/2)*s(\hat{\alpha}),\hat{\alpha}+t(n-2;1-\alpha/2)*s(\hat{\alpha})],\quad s^2(\hat{\alpha})=MSE(\frac{1}{n}+\frac{(\bar{X})^2}{\sum(X_i-\bar{X})^2})$$

②가설 검정 $H_0 : \alpha = \alpha_0 = 0$: 검정 통계량 $T = \frac{\hat{\alpha} - \alpha_0}{s(\hat{\alpha})} = \frac{\hat{\alpha} - 0}{s(\hat{\alpha})}$의 절대값이 $t(n-2;1-\alpha/2)$ 보다 크면 귀무가설을 기각하고 적으면 귀무가설을 채택한다.

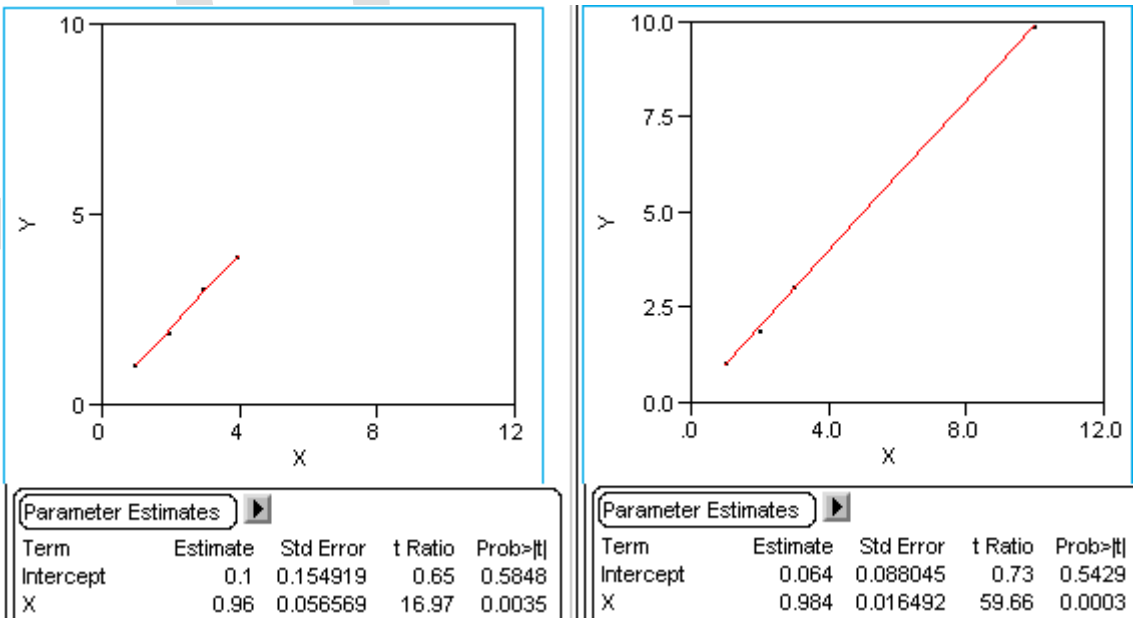
2.4.3 회귀계수 추론에 대한 Comment

Abnormality (비정규성)

오차항 e_i가 정규 분포를 따르지 않으면 y_i가 정규 분포를 따르지 않게 되고 회귀계수 추정치 $\hat{\alpha}, \hat{\beta}$ (y_i의 함수)도 정규 분포를 따르지 않게 된다. 즉 더 이상 회귀계수 가설 검정에 있어서 t-분포를 사용할 수 없게 된다. 그러나 안심하자. 비록 y_i가 정규 분포를 따르지 않더라도 표본 개수 n이 증가하면 $\hat{\alpha}, \hat{\beta}$는 근사적으로 정규 분포에 근사한다. 그러므로 t-분포를 사용하여 가설 검정할 수 있다. 회귀분석에서 비정규성 문제는 심각한 것이 아니다.

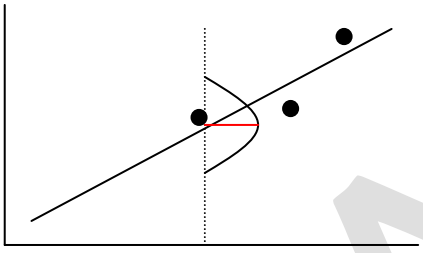
설명변수 X 값의 간격

설명변수 X의 간격이 넓어질수록 $\sum(X_i - \bar{X})^2$이 커지므로 $s(\hat{\alpha}), s(\hat{\beta})$는 줄어들어 t-값은 커지고 (F-값도 커진다) 회귀계수가 유의할 가능성이 높아진다. 다음은 4개의 자료가 측정되었는데 X가 등간격일 경우와(왼쪽) 하나가 다른 관측치에 비해 멀리 떨어진 경우(오른쪽) 차이점을 살펴보자. 기울기 회귀계수 추정치($\hat{\beta}$) 값은 비슷하지만 추정 오차의 차이로 인하여 t-검정 통계량이 크게 차이가 난다. 정말로 오른쪽 경우가 더 유의한가? 그렇지 않다. 그러므로 X 값이 다른 관측치에 비해 너무 멀리 떨어진 경우에는 그 관측치를 회귀 분석에서 제외하기 바란다. 또한 수집된 설명변수 데이터 범위를 많이 벗어나는 곳에서는 예측치를 구하지 말자.



2.4.4 $\hat{E}(Y_0)$에 대한 추론

임의의 설명변수 값에 대해 종속변수의 평균을 추정해 보자. $E(y_0)=\mu_{y|x_0}=\alpha+\beta x_0$. 예를 들면 광고비를 500만$ 쓰면 고객 상품인지도의 평균은 얼마일까? 관심을 갖는 설명 변수의 값을 x_0라 하면 $E(y_i)=\alpha+\beta x_0$이고 이것에 대한 점추정치는 $\hat{E}(y_0)=\hat{\mu}_{y|x_0}=\hat{\alpha}+\hat{\beta}x_0$ 이다.



x_0=500만

$\frac{\hat{E}(Y_0)-E(\hat{E}(Y_0))}{s\{\hat{E}(Y_0)\}}$의 분포 함수

$\hat{E}(Y_0)$는 추정치 $\hat{\alpha},\hat{\beta}$의 선형 결합 함수이므로 정규분포 따른다. ($\because \hat{\alpha}\sim Normal\ \ \hat{\beta}\sim normal$)

$\hat{E}(Y_0)$의 평균 $E(\hat{E}(Y_0))=\alpha+\beta X_0$, 분산 $V(\hat{E}(Y_0))=\sigma^2(\hat{E}(Y_0))=\sigma^2[\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum(X_i-\bar{X})^2}]$ 추정치 $\hat{E}(Y_0)$ 분산의 추정치 $s^2(\hat{E}(Y_0))=MSE[\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum(X_i-\bar{X})^2}]$이다. 추정치 $\hat{\alpha},\hat{\beta}$의 분포 유도와 동일한 방법으로 분포 함수를 유도하면 $\frac{\hat{E}(Y_0)-E(Y_0)}{s\{\hat{E}(Y_0)\}}\sim t(n-2)$, $s\{\hat{E}(Y_0)\}=\sqrt{MSE[\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum(X_i-\bar{X})^2}]}$

$E(Y_0)$에 대한 신뢰구간과 가설 검정

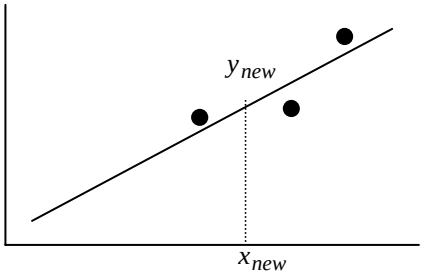
①100(1-α)% 신뢰구간 $\hat{\alpha}+\hat{\beta}x_0-t(n-2;1-\alpha/2)*s(\hat{E}(Y_0)),\hat{\alpha}+\hat{\beta}x_h+t(n-2;1-\alpha/2)*s(\hat{E}(Y_0))$

②유의수준 α 가설 검정 $H_0:E(Y_0)=\mu_{y|x_0}$

$T=\frac{\hat{E}(Y_0)-\mu_{y|x_0}}{s\{\hat{E}(Y_0)\}}$의 절대값이 $t(n-1;1-\alpha/2)$보다 크면(양측 검정) 귀무가설을 기각한다.

2.4.5 새로운 관측치 Y_{new} 에 대한 추론

설명변수의 값이 주어졌을 때 관측되는 종속변수의 값을 추정해 보자. 주어진 설명변수의 값을 x_{new}라 하고 이에 대응하는 종속변수의 값을 y_{new}라 하자.



점 추정치는 $\hat{y}_{new} = \hat{\alpha} + \hat{\beta}x_{new}$ 이다. 이것은 $E(y_0)$ 의 추정치와 동일하다. 차이가 있다면 새로운 관측치 Y_{new} 의 점추정치의 분산이 σ^2 만큼 크다는 것이다. 수집된 데이터에 없는 설명변수에 대한 종속변수 값의 예측치은 이 방법을 사용하고 데이터에 있는 설명변수 값에 대한 추정치는 2.4.4절(종속변수 평균에 대한 관측치) 방법을 사용하자.

설명변수 새로운 값에 대한 종속변수 예측치를 구할 때 관측된 설명변수 범위 내의 설명변수 값들에 대해서만 한정하기를 강력 권한다. (2.4.3절에서 설명)

$\frac{\hat{Y}_{new} - E(\hat{Y}_{new})}{s\{\hat{Y}_{new}\}}$ 의 분포

$\hat{Y}_{new}$도 추정치 $\hat{\alpha}, \hat{\beta}$의 선형 결합 함수이므로 정규분포를 따른다. ($\because \hat{\alpha} \sim Normal \quad \hat{\beta} \sim normal$)

평균 $E(\hat{y}_{new}) = \alpha + \beta x_{new}$ 이고 분산은 $\sigma^2\{\hat{Y}_{new}\} = \sigma^2[1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{\sum(x_i - \bar{x})}]$ 이다. $E(\hat{Y}_0)$ 분산보다 σ^2 만큼 크다. 다음이 성립한다.

$$\frac{\hat{Y}_{new} - E(\hat{Y}_{new})}{s\{\hat{Y}_{new}\}} \sim t(n-2) \quad , \quad s\{\hat{Y}_{new}\} = MSE[1 + \frac{1}{n} + \frac{(X_{new} - \bar{X})^2}{\sum(X_i - \bar{X})}]$$

Y_{new} 에 대한 신뢰구간과 가설 검정

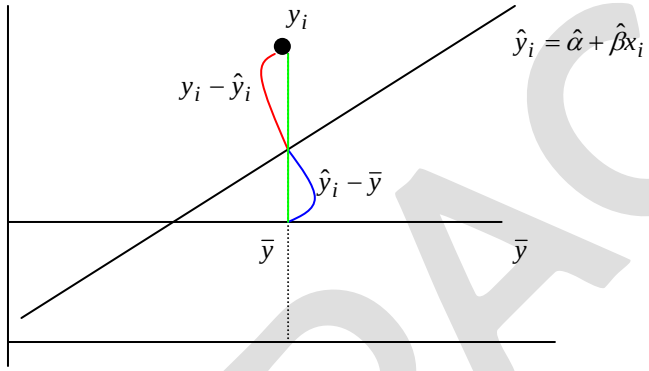
①100(1-α)% 신뢰구간 $\hat{\alpha} + \hat{\beta}x_h - t(n-2;1-\alpha/2) * s(\hat{Y}_{new}), \hat{\alpha} + \hat{\beta}x_h + t(n-2;1-\alpha/2) * s(\hat{Y}_{new})$,

②유의수준 α 가설 검정 $H_0 : Y_{new} = Y_h \quad T = \frac{\hat{Y}_{new} - Y_{new}}{s\{\hat{Y}_{new}\}}$ 의 절대값이 $t(n-1;1-\alpha/2)$ 보다 크면 (양측 검정) 귀무가설을 기각한다.

2.5 회귀분석에 분산분석적 접근

회귀 분석을 분산분석적 측면에서 다루는 것은 단순 회귀에서는 새로운 것이 없으나(단순 회귀 모형에서는 회귀계수에 대한 t-검정은 분산분석의 F-검정과 동일하다. $F(1,n-)=t^2(n-2)$) 보다 복잡한 회귀 모형을 다루는데 도움을 얻을 것이다.

2.5.1 변동 분할



분산분석 접근은 종속변수 Y에 관련된 총변동과 자유도 분할에 근거한다. 총변동(SSTO, SST, Total Sum of Square)은 종속 변수의 관측치와 평균의 편차(deviation) $(y_i - \bar{y})$ 제곱합을 의미하며 이는 종속 변수가 가진 정보이다.

$$\text{총변동 } SSTO = \sum (y_i - \bar{y})^2 \quad \text{(초록색 부분) } SST = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n} = \sum Y_i^2 - n\bar{Y}^2$$

회귀 모형에서 데이터에 포함된 불확실성(uncertainty)은 적합 회귀선(fitted regression line, 추정 회귀식)으로부터 관측치가 얼마나 벗어나 있나를 의미하며 이것에 대한 측정은 $(y_i - \hat{y}_i)$ 이고 제곱 합을 오차변동(Error Sum of Squares, SSE) 혹은 오차 제곱합(자승합)이라 하며 적합 회귀식에 의해 설명되지 않는 변동에 해당된다. 이 오차 변동을 $(n-2)$ 로 나눈 값을 MSE 라 하면 이는 오차의 분산에 대한 추정치로 사용한다.

$$\text{오차 변동: } SSE = \sum (y_i - \hat{y}_i)^2 \quad \text{(빨간 부분)}$$

두 변동의 차이를 회귀변동(Regression Sum of Squares, SSR) 혹은 모형 변동(Model SS)이라 하며 적합된 회귀식이 데이터의 관계를 얼마나 잘 설명하는지 나타낸다.

$$\text{회귀(모형)변동: } SSR = \sum (\hat{y}_i - \bar{y}_i)^2 \text{ (파란부분)}$$

$$SSR = \hat{\beta}(\sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n}) = \hat{\beta}^2 \sum (X_i - \bar{X})^2$$

2.5.2 자유도 분할

총변동의 자유도(관측치 중 자유로운 개수, 관측치 하나 하나는 독립적이고 정보를 갖고 있다)는 평균이 하나($\bar{y}$) 추정되었으므로 $(n-1)$이다. SSE의 자유도는 $(n-2)$이다. 왜냐하면 $\hat{\alpha},\hat{\beta}$가 두 개 추정되었기 때문이다. SSR의 자유도는 SST 자유도로부터 SSE 자유도를 뺀 값으로 1이다. 회귀모형에서 총변동의 자유도는 $(n-1)$ 이고 모형변동 자유도는 설명변수의 개수 p, 오차변동의 자유도는 $(n-p-1)$ 이다.

2.5.3 평균 변동과 (평균 제곱합) 기대 평균 변동

변동 합(제곱합)을 자유도로 나눈 값을 평균 변동이라 한다.

$$MSR = \frac{SSR}{1} \text{ (Mean Sum of squares of Regression 회귀 평균변동)}$$

$$MSE = \frac{SSE}{(n-2)} \text{ (Mean Sum of squares of Error 오차 평균변동)}$$

EMS (Expected Mean Squares 기대 평균변동)

$$E(MSE) = \sigma^2, \quad E(MSR) = \sigma^2 + \beta^2 \sum (X_i - \bar{X})^2 \text{ 이다.}$$

이에 대한 증명은 다음과 같다.

$$\frac{SSE}{\sigma^2} = \frac{\sum (Y_i - \hat{Y}_i)^2}{\sigma^2} = \frac{\sum (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2}{\sigma^2} \sim \chi^2(n-2) \rightarrow E(MSE) = \sigma^2$$

$$SSR = \hat{\beta}^2 \sum (X_i - \bar{X})^2 \text{ 와 } V(\hat{\beta}) = \sigma^2(\hat{\beta}) = \frac{\sigma^2}{\sum (X_i - \bar{X})^2} \rightarrow E(MSR) = \sigma^2 + \beta^2 \sum (X_i - \bar{X})^2$$

MSE의 기대값인 σ^2 이므로 MSE는 σ^2 의 불편 추정치이다. MSR 기대값의 의미는? 만약 $\beta = 0$ (설명변수가 유의하지 않음) 혹은 모든 관측치(X_i)가 평균($\bar{X}$)과 같으면 EMR은 σ^2 이므로 F-값은 1이다. 실제 현실에서는 모든 관측치(X_i)가 평균($\bar{X}$)과 같을 경우는 발생하지 않으므로 설명변수가 유의하지 않으면 MSR은 σ^2 에 근사하고 F-값은 1에 근사한다. 즉 F-값이 커져야 설명변수는 유의하다.

2.5.4 F-검정

$H_0 : \beta = 0$ 이면 오차변동과 모형변동은 같아지므로 다음 통계량에 의해 $H_0 : \beta = 0$ 의 유의성을 검정할 수 있을 것이다. 그러므로 $F^* = \frac{MSR}{MSE}$ 의 값이 커지면 귀무가설 $H_0 : \beta = 0$ 을 기각할 가능성이 높아지게 된다. 귀무가설($H_0 : \beta = 0$)하에서는 SSR/σ^2 과 SSE/σ^2 가 서로 독립임을 이용하면 $F^* = \frac{MSR}{MSE} = \frac{\frac{SSR}{\sigma^2}/1}{\frac{SSE}{\sigma^2}/(n-2)} = \frac{\chi^2(1)/1}{\chi^2(n-2)/(n-2)} \sim F(1, n-2)$ 이 성립한다. 그러므로 $F^* \le F(1-\alpha; 1, n-2)$ 이면 귀무가설 $H_0 : \beta = 0$ (설명 변수는 종속 변수에 영향을 미치지 않는다) 채택하고 $F^* > F(1-\alpha; 1, n-2)$ 이면 귀무가설을 기각한다.

F-검정의 모형에 설정한 설명변수 전체에 대한 유의성 검정에 사용된다. 즉 F-검정은 $H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$ (회귀모형에 고려된 설명변수 모두는 유의하지 않다)의 유의성을 검정한다. 그러므로 분산분석적 측면에서 모형 변동($SSR = \sum (y_i - \bar{y}_i)^2$ 은 고려된 설명변수들의 유의성 검정이다. F-검정 결과 귀무가설이 기각되면 "설명변수 중 적어도 하나는 유의함을 알 수 있다" ⇔ "$\beta_k \neq 0$, at least one k ". 그러므로 모형 변동($SSR = \sum (y_i - \bar{y}_i)^2$ 은 고려된 모형의 설명변수에 의한 설명력의 척도이다.

F-검정 결과 유의하지 않으면 유의한 설명변수가 없다는 의미이므로 더 이상의 분석은 의미가 없다. 유의하면 설명변수 각각에 대한 유의성 검정인 t-검정을 실시하면 된다.

t-검정과의 관계

$SSR = \beta^2 \sum (X_i - \bar{X})^2$ 이고 $s^2(\hat{\beta}) = \frac{MSE}{\sum (X_i - \bar{X})^2}$ 이므로 다음이 성립하므로 **단순회귀분석**에서는 분산분석의 F-검정과 기울기 회귀계수에 대한 t-검정은 동일하다. 단순회귀에서는 모형에 대한 F-검정이나 설명변수(회귀계수)에 대한 t-검정은 동일하다.

$$F^*(1, n-2) = \frac{\beta^2 \sum (X_i - \bar{X})^2}{MSE} = \frac{\hat{\beta}^2}{s^2(\hat{\beta})} = (\frac{\hat{\beta}}{s(\hat{\beta})})^2 = t^2(n-2)$$

2.5.5 분산분석표

변동 (source)	SS(자승합)	Df (자유도)	MS (평균 자승합)	EMS (기대 평균 자승합)
Regression (모형, 회귀)	$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$	$p = 1$	$MSR = SSR / p$	$E(MSR) = \sigma^2 + \beta^2 \sum (X_i - \bar{X})^2$
Error (오차)	$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$n - p - 1 = n - 2$	$MSE = SSE/(n-2)$	$E(MSE) = \sigma^2$
Total (총변동)	$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$	$n - 1$		$F = \frac{MSR}{MSE} \sim F(1, n-1)$

다중 회귀모형($y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi} + e_i$)에서는 $H_0 : \beta_1 = \beta_1 = ... = \beta_p = 0$ (설명변수가 모두 유의하지 않다) 유의성 검정은 F-검정을 실시하고 각 설명변수에 대한 유의성 검정은 t-검정을 실시한다.

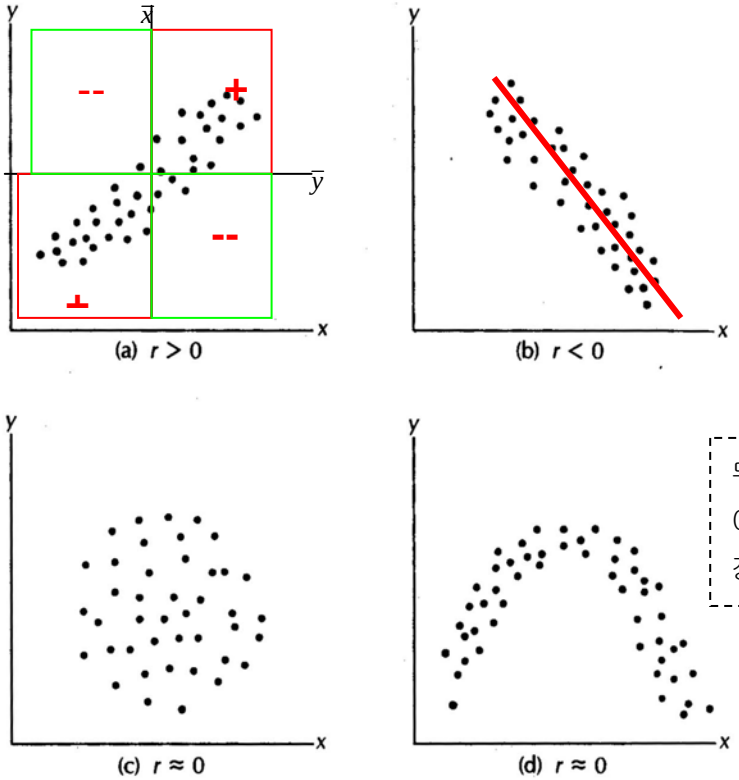
2.5 상관분석

두 변수간의 (선형) 관계를 분석하는 방법으로 상관 정도는 상관 계수에 의해 측정하며 상관계수에 대한 검정은 t-검정을 이용한다.

2.5.1 상관 계수

상관계수는 두 변수 간의 **선형**(직선) 관계가 존재하는 알아보는 방법이다. 회귀 분석과 유사하지만 인과 관계에 대한 분석은 아니다. 상관계수는 다음과 같이 구한다. 이를 Pearson 상관계수(correlation coefficient)라 한다.

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

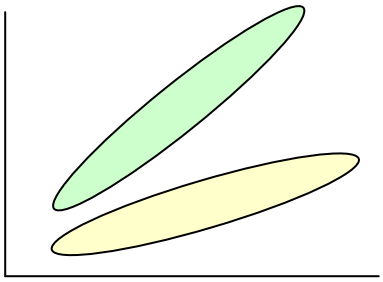


두 변수간의 관계는 2차식이나 선형관계 정도를 측정하는 상관계수는 0이다.

점들이 직선에 모여 있을수록 상관계수는 커진다. 상관 계수의 부호가 양이면 한 변수의 값이 커질수록(작아질수록) 다른 변수의 값도 커짐(작아짐)을 의미하며 음이면 한 변수의 값이 커질수록(작아질수록) 다른 변수의 값도 작아짐(커짐)을 의미한다.

표본의 크기가 커지면 상관계수 값이 커지므로 상관계수 값이 얼마 이상이어야 기준은 없으므로 가설검정에 의한 유의확률을 계산하기 바란다. 상관계수에 대해 다음 사항을 주의하기 바란다.

- 상관계수는 두 변수간의 선형 관계를 알아보는 것이다. 이차 관계의 상관계수는 0이다.
- 상관계수는 점들이 직선에 모여 있는 정도를 나타내는 지표이지 직선의 기울기의 크기를 나타내는 것은 아니다. 오른쪽 그림에서 두 타원의 상관계수는 동일하다.



Pearson 상관계수는 측정형 변수 간의 상관 정도를 나타낸다. 데이터가 순서형이거나 가질 수 있는 값이 10개 이하인 경우 (예:리커트 Likert 척도) 비모수적인 방법으로 상관계수를 구하는 것이 좋다. Spearman 순위(rank order) 상관계수와 Kendall의 τ 이 대표복인 비모수적 방법이다.

$$r_s = \frac{\sum (R_x - \overline{R}_x)(R_y - \overline{R}_y)}{\sqrt{\sum (R_x - \overline{R}_x)^2}\sqrt{\sum (R_y - \overline{R}_y)^2}} \approx 1 - \frac{6\sum d_i^2}{n(n^2-1)}, \; R \text{은 관측치 순위}, \; d_i = R_{x_i} - R_{y_i} \text{이다.}$$

$$\tau = \frac{\sum_{i<j} sign(x_i - x_i)sign(y_i - y_j)}{\sqrt{(T_0 - T_x)(T_0 - T_y)}}, \; sign(w) = \begin{cases} -1, & w < 0 \\ 0, & w = 0 \\ 1, & w > 1 \end{cases}, \; T_0 = n(n-1)/2, \; T_x = \sum t_i(t_i - 1)/2,$$

t_i 는 동일한 x_i 의 i-번째 그룹 내의 관측치 개수이다.

2.5.2 상관계수 추론

귀무가설: $H_0 : \rho = 0$ (모집단 상관계수는 0이다. 두 변수는 서로 독립이다)

대립가설: $H_a : \rho \neq 0$

$$\text{검정통계량: } T = \frac{r}{\sqrt{(1-r^2)/(n-2)}} \sim t(n-2)$$

만약 귀무가설이 $H_0 : \rho = \rho_0 \neq 0$ (예: 모집단의 상관계수가 0.7이다.)이라면 다음 절차를 이용하여 상관 관계에 대한 가설을 검정한다.

$$\text{검정통계량: } T = 0.5 \ln \frac{1+r}{1-r} \sim N(0.5 \ln \frac{1+\rho_0}{1-\rho_0}, \frac{1}{n-3})$$

위 사실을 이용하여 두 모집단 상관계수 차이 검정을 다음 절차에 의해 실시할 수 있다.

$$z(x) = 0.5 \ln \frac{1+r_x}{1-r_x}, z(y) = 0.5 \ln \frac{1+r_y}{1-r_y}$$

$$z = \frac{z(x) - z(y)}{\sqrt{1/(n_x - 3) + 1/(n_y - 3)}} \sim N(0,1)$$

2.5.3 회귀계수와 관계

$Sxx = \sum (x_i - \overline{x})^2, Syy = \sum (y_i - \overline{y})^2$ 이라 하면 회귀 모형에서 기울기 회귀 계수 추정치와 상관 계수는 관계는 $\hat{\beta} = \sqrt{\frac{Syy}{Sxx}} r$ 이다. 그러므로 다음 사실을 알 수 있다.

- 기울기의 부호와 상관 계수의 부호는 같다.
- 단순 회귀분석 기울기에 대하 유의성 검정과 상관계수 유의성 검정은 동일하다.

• 타원이 좁을수록 상관계수는 커지면 회귀분석의 정도(precision)은 높아진다. 다음 장에서 설명하게 될 회귀분석의 결정계수(R^2)는 상관계수는 제곱과 동일하다.

2.6 원점을 지나는 회귀 직선

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. 그러나 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. 예를 들어 비용과 생산량과의 관계를 보거나 약 복용량에 따른 바이러스 감소량의 관계를 보는 경우 원점을 지나는 회귀 모형을 고려하면 된다. SAS에서는 NOINT 옵션 사용하면 된다.

2.7 결정계수

회귀계수 추정과 검정, 종속변수 관측치에 대한 예측치($\hat{Y}_{new}$), 평균 예측치($\hat{E(Y_0)}$)에 대해 살펴 보았으나 두 변수 간의 선형관계 정도를 나타낸 통계량은 없었다. 이에 다음과 같이 결정계수(Coefficient of Determination)를 정의한다. $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$ 결정계수는 두 변수 간의 선형 관계 정도가 높으면 (관측치들이 직선 가까이에 모여 있다는 것을 의미) 결정계수는 1에 가까워진다. 특히 단순 회귀 모형에서는 상관계수는 $r = \pm\sqrt{R^2}$ 이 성립한다.

(참고: $\beta = \sqrt{\frac{S_{yy}}{S_{xx}}} r$) 결정계수는 단순히 선형관계 정도를 나타내는 수치일 뿐 검정할 수 있는 검정통계량이 존재하지 않아 단지 지표로 사용될 뿐이다. 특히 설명 변수가 이산형(설문지 Likert 척도)인 경우 매우 낮아지는 경향이 있고 관측치가 많아지면 커지는 경향이 있어 선형관계 정도를 나타내는 좋은 지표는 아니다.

유의하지 않은 설명변수라도 모형에 삽입되면 결정계수 값은 올라가므로 모형의 유의성 비교에는 사용하지 않는다. 대신 설명변수의 개수에 의해 조정된 수정(adjusted) 결정계수를 사용한다. 이것은 다중회귀에서 상세히 다루기로 한다.

2.8 통계소프트웨어 사용하기

EXAMPLE 2-5 단순회귀 하기

AD.xls (엑셀 데이터)

1983년 미국 21개 기업 광고비(SPEND, 단위: 백만$)가 소비자 평가도(RATE)를 조사한 것이다. 회귀분석을 다음과 같이 실시하시오.

(1)산점도를 그리시오. 2.1.2절 (페이지 22 참고)

(2)광고비의 회귀계수 OLS 추정치와 추정분산을 구하시오.

(3)광고비의 유의성을 검정하시오.

(4)유의하다면 회귀모형을 적고 해석하시오. 회귀계수의 95% 신뢰구간도 구하자.

(5)유의하다면 광고비가 40.1인 경우 평가도 예측치를 구하시오. 95%신뢰구간

(6)유의하다면 광고비가 50인 경우 평가도 예측치를 구하시오. 95%신뢰구간

(1) 풀이

2.1.2절의 산점도(페이지 22) 결과 광고비 100 이상인 기업(3개, MacDonald, Ford, AT&T)은 기업 구조 상 광고비를 많이 편성하는 기업으로 이를 제외하는 것이 적절하다. 그러므로 향후 회귀분석에서는 이를 제외한 17개 기업만을 사용하기로 한다.

SAS

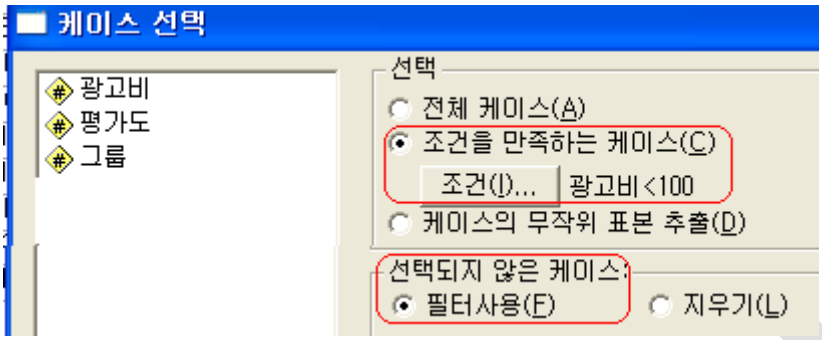
```
data ad0;
    set ad;
        if (spend>100) then delete;
run;
```

SPSS

	기업	광고비	평가도	그룹
1	PEPSI	74.10	99.60	1.00
2	STROH'S	19.30	11.70	2.00
3	FED'L EX	22.90	21.90	2.00
4	BURGER K	82.40	60.80	3.00
5	COCO-COL	40.10	78.60	1.00
6	MC DONAL	185.90	92.40	3.00

데이터(D) ▶ 케이스 정렬(O)...을 선택한 후 "광고비" 변수로 정렬한 후 지우려는 개체의 행에서 지우거나 다음 방법을 이용한다.

데이터(D) ▶ 케이스 선택(C)...을 선택하고 아래와 같이 화면을 설정한다.

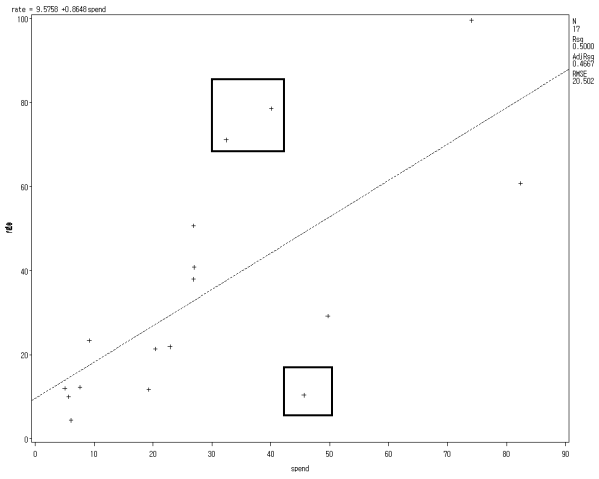


	기업	광고비	평가도	그룹	filter_$
1	PEPSI	74.10	99.60	1.00	1
2	STROH'S	19.30	11.70	2.00	1
3	FED'L EX	22.90	21.90	2.00	1
4	BURGER K	82.40	60.80	3.00	1
5	COCO-COL	40.10	78.60	1.00	1
6	MC DONAL	185.90	92.40	3.00	0
7	MCI	26.90	50.70	2.00	1
8	DIET COL	20.40	21.40	1.00	1

(2)-(3) 풀이

SAS

```
proc reg data=ad0;
      model rate=spend;
      plot rate*spend;
run;
```



여전히 이상치가 존재하는 것 같다.

이상치를 판단하는 검정통계량을 배울 때까지 잠시 덮어 두자.

분산분석표: 유의확률이 0.0015이므로 회귀모형은 유의하다. 개별 설명변수의 유의성 검정은 아래 t-검정 이용하면 된다. 데이터가 17(3개는 이상치로 제외)개이므로 총변동의 자유도는 16이다. F-검정 유의확률이 0.0015이므로 회귀모형에 설정한 설명변수 광고비는 유의하다. 즉 광고비는 소비자 평가도에 영향을 미침을 알 수 있다.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6305.53455	6305.53455	15.00	0.0015
Error	15	6304.93604	420.32907		
Corrected Total	16	12610			

Root MSE = $\sqrt{MSE} = \sqrt{420.32}$, R-square=결정계수, Dependent Mean은 종속변수(평가도)의 평균, Adj-R-Sq는 $1 - \frac{SSE/(n-p-1)}{SST/(n-1)} = 1 - \frac{420.32}{12610/16} = 0.4667$ (결정계수의 문제점을 보완하기 위한 통계량) Coeff. Var은 종속변수의 변동계수($s_y / \bar{y} * 100(\%)$, 분산의 비교에 사용)이다.

Root MSE	20.50193	R-Square	0.5000
Dependent Mean	35.07647	Adj R-Sq	0.4667
Coeff Var	58.44923		

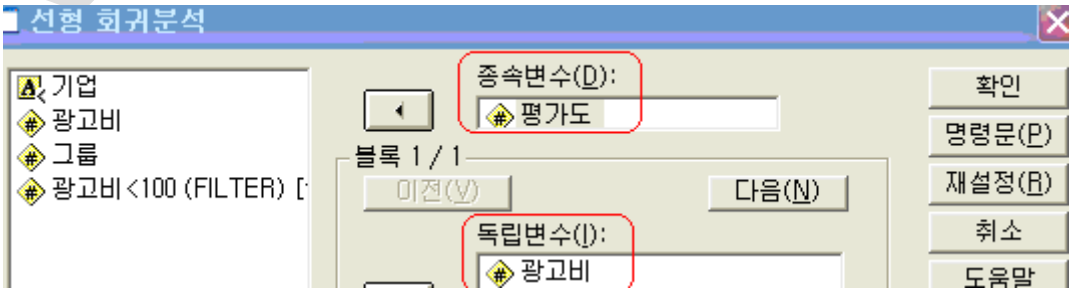
광고비 회귀계수의 유의확률이 0.0015로 유의수준 0.05보다 작으므로 유의하다고 할 수 있다. 회귀계수가 0.86이므로 광고를 많이 할수록 평가가 높아짐을 알 수 있다. 광고비를 단위 1만큼 더 사용하면 평가는 0.86만큼 증가한다.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	9.57579	8.25066	1.16	0.2639
spend	1	0.86477	0.22327	3.87	0.0015

최종 회귀모형: $\begin{matrix} 평가 = 9.57 + 0.86 * 광고비 \\ (t = 3.97,\ p = 0.0015) \end{matrix}$

SPSS 분석(A) ▶ 회귀분석(R) 선형(L)... 메뉴를 선택하고 아래와 같이 메뉴를 설정한다. 일단은 설명변수와 종속변수만 지정하자. 나머지는 default 사용하자. 산점도 그리기는 페이지 23을 참고하기 바란다.



SAS 결과와 동일하다. 하나 더 출력되는 것이 있다면 표준화 회귀계수(standardized beta coefficient)이다. 이는 설명변수를 표준화(standardization)하여 얻은 회귀계수이다. 종속변수에 대한 설명변수간의 영향력을 비교하는데 사용한다. 자세한 내용은 다중 회귀에서 다루기로 한다.

모형 요약

모형	R	R 제곱	수정된 R 제곱	추정값의 표준오차
1	,707[a]	,500	,467	20,50193

분산분석[b]

모형		제곱합	자유도	평균제곱	F	유의확률
1	선형회귀분석	6305,535	1	6305,535	15,001	,002[a]
	잔차	6304,936	15	420,329		
	합계	12610,471	16			

계수[a]

모형		비표준화 계수		표준화 계수	t	유의확률
		B	표준오차	베타		
1	(상수)	9,576	8,251		1,161	,264
	광고비	,865	,223	,707	3,873	,002

(4)회귀계수 신뢰구간 구하기

SAS

```
proc reg data=ad0;
    model rate=spend/clb alpha=0.05;
run;
```

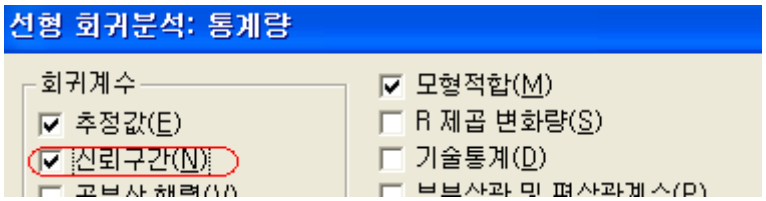
회귀계수 신뢰구간을 구하려면 CLB(Confidence Limit for Beta) 옵션을 사용하면 된다. 신뢰수준 0.95이면 $\alpha = 0.05$ (default) 사용하면 한다. 90% 신뢰구간이면 $\alpha = 0.1$ 사용한다.

기울기(설명변수 spend)의 95% 신뢰구간은 0을 포함하고 있지 않으므로 유의하다. 검정결과와 동일하다. 사실 회귀계수 신뢰구간은 별로 사용하지는 않는다.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	9.57579	8.25066	1.16	0.2639	-8.01008	27.16166
spend	1	0.86477	0.22327	3.87	0.0015	0.38888	1.34067

SPSS 선형 회귀 모형 설정 창에서 "통계량" 옵션의 신뢰구간을 선택하면 된다.



(5)-(6)

수집된 데이터에 없는 설명변수의 값에 대해 예측치나 예측치 신뢰구간을 구하려면 데이터 제일 마지막 부분에 설명변수 값과 종속변수는 결측치(.)으로 하여 데이터를 입력한다. 물론 마지막 관측치는 회귀모형에 사용되지 않는다.

SAS

```
KIBBLES     6.10    4.40    3.00
New         50      .       .
run;

proc reg data=ad0;
    model rate=spend/clm cli p r alpha=0.05;
run;
```

○P(predicted) 종속변수 예측치 $\hat{y}_i$ ○R(residual) 잔차($r_i = y_i - \hat{y}_i$)

○CLM(confidence limit for mean) $E(y_0)$ 의 신뢰구간

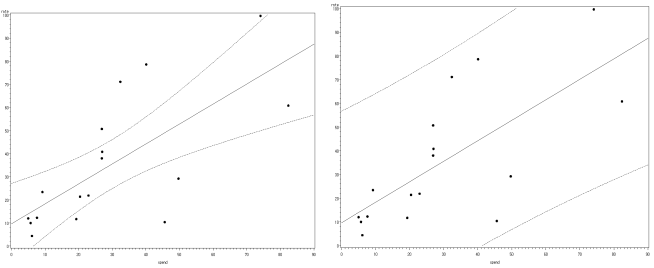
○CLI(confidence limit for individual) y_{new} 의 신뢰구간

SAS에는 종속변수 개별 예측치 신뢰구간, 평균 예측치 신뢰구간을 그릴 수 있는 PROC가 있다. RL의 의미는 Regression Line의 약어이다.

$E(Y_0)$ 와 Y_{new} 의 신뢰구간 그리기

```
proc gplot data=ad0;
    symbol i=rlclm v=dot;
    plot rate*spend;
run;

proc gplot data=ad0;
    symbol i=rlcli v=dot;
    plot rate*spend;
run;
```



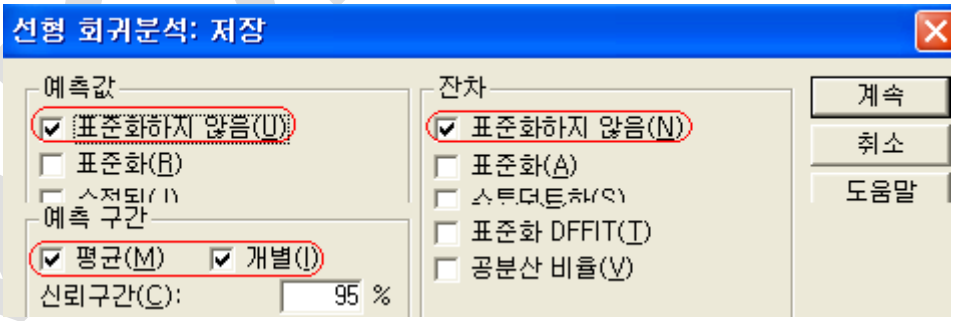
언급한대로 평균에 대한 예측구간, 개별 관측치에 대한 신뢰구간에 비해 좁다. 잔차는 관측치와 예측치 $\hat{y}_i$의 차이이다. $\hat{y}_i$는 최종 회귀모형 평가 = 9.57 + 0.86 * 광고비에 의해 계산된 값이다. 회귀모형 추정결과를 비교해 보라. 마지막 관측치 없는 결과와 동일하다.

Obs	Dep Var rate	Predicted Value	95% CL Mean		95% CL Predict		Residual	Std Error Residual	Student Residual
1	99.6000	73.6556	49.9266	97.3846	23.9298	123.3814	25.9444	17.216	1.507
2	11.7000	26.2659	14.6110	37.9209	-18.9604	71.4923	-14.5659	19.759	-0.737
3	21.9000	29.3791	18.3266	40.4317	-15.6958	74.4540	-7.4791	19.835	-0.377
4	60.8000	80.8332	53.5132	108.1533	29.2971	132.3693	-20.0332	16.001	-1.252
5	78.6000	44.2533	32.5131	55.9934	-0.9952	89.5017	34.3467	19.748	1.739
6	50.7000	32.8382	22.1684	43.5081	-12.1444	77.8208	17.8618	19.881	0.898
7	21.4000	27.2172	15.7702	38.6642	-17.9561	72.3904	-5.8172	19.786	-0.294
8	40.8000	32.9247	22.2602	43.5892	-12.0566	77.9060	7.8753	19.882	0.396
9	10.4000	49.0095	35.9282	62.0908	3.3947	94.6243	-38.6095	19.562	-1.974
10	12.0000	13.8997	-1.8528	29.6522	-32.5517	60.3510	-1.8997	19.124	-0.0993
11	29.2000	52.5551	38.2426	66.8676	6.5721	98.5381	-23.3551	19.371	-1.206
12	38.0000	32.8382	22.1684	43.5081	-12.1444	77.8208	5.1618	19.881	0.260
13	10.0000	14.5050	-1.0026	30.0127	-31.8639	60.8739	-4.5050	19.168	-0.235
14	12.3000	16.1481	1.2876	31.0085	-30.0084	62.3046	-3.8481	19.280	-0.200
15	23.4000	17.5317	3.1947	31.8687	-28.4589	63.5223	5.8683	19.367	0.303
16	71.1000	37.5945	26.9058	48.2832	-7.3926	82.5816	33.5055	19.879	1.685
17	4.4000	14.8509	-0.5183	30.2202	-31.4719	61.1737	-10.4509	19.192	-0.545
18	.	52.8145	38.4057	67.2234	6.8015	98.8276	.	.	.

SPSS 데이터 창 마지막에 설명변수 데이터를 입력한다.

19	CREST	32.40	71.10	2.00	1
20	KIBBLES	6.10	4.40	3.00	1
21	NEW	50.00	.	.	1

선형 회귀분석 창의 "저장" 옵션을 아래와 같이 선택하면 된다. "표준화하지 않음"은 예측치 $\hat{y}_i$, 예측구간은 $E(Y_0)$와 Y_{new} 값, "표준화하지 않음" 잔차는 r_i을 의미한다.



	기업	광고비	평가도	그룹	filter_$	PRE_1	RES_1	LMCI_1	UMCI_1	LICI_1	UICI_1
1	PEPSI	74.10	99.60	1.00	1	73.656	25.944	49.927	97.385	23.930	123.38
2	STROH'S	19.30	11.70	2.00	1	26.266	-14.57	14.611	37.921	-18.96	71.492
3	FED'L EX	22.90	21.90	2.00	1	29.379	-7.479	18.327	40.432	-15.70	74.454
4	BURGER K	82.40	60.80	3.00	1	80.833	-20.03	53.513	108.15	29.297	132.37
5	COCO-COL	40.10	78.60	1.00	1	44.253	34.347	32.513	55.993	-.9952	89.502
6	MC DONAL	185.90	92.40	3.00	0	.	.	.	.	.	.
7	MCI	26.90	50.70	2.00	1	32.838	17.862	22.168	43.508	-12.14	77.821
20	KIBBLES	6.10	4.40	3.00	1	14.851	-10.45	-.5183	30.220	-31.47	61.174
21	NEW	50.00	.	.	1	52.815	.	38.406	67.223	6.8015	98.828

상관계수

단순회귀모형의 회귀계수와 상관계수의 관계를 알아보기 위하여 상관계수를 구해보자.

SAS

이상치로 보이는 3개를 제외한 "AD0" 데이터를 이용하여 상관분석을 실시하였다. DATA 문 옆에 아무 옵션을 사용하지 않으면 Pearson 상관계수가 출력된다. 만약 비모수 상관계수를 출력하기 원하면 KENDALL 혹은 SPEARMAN이라 적어 주면 된다.

```
proc corr data=ad0;
    var rate spend;
run;
```

유의확률이 0.0015이므로 평가도와 광고비의 상관관계는 유의하다. 상관계수는 0.707이므로 광고비가 높아지면 평가도가 높아진다.

피어슨 상관 계수, N = 17
H0: Rho=0 검정에 대한 Prob > |r|

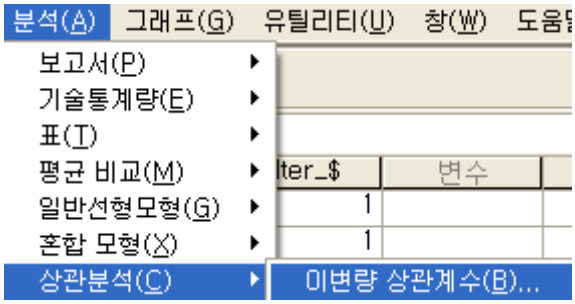
	rate	spend
rate	1.00000	0.70712 0.0015
spend	0.70712 0.0015	1.00000

페이지 40의 상관계수와 회귀계수의 관계식이 맞는지 숫자로 살펴보자. 다음은 SAS에 출력된 각 변수의 기초통계량이다. 만약 이것을 출력하지 않으려면 PROC CORR 문장에 옵션으로 NOSIMPLE을 사용하면 된다.

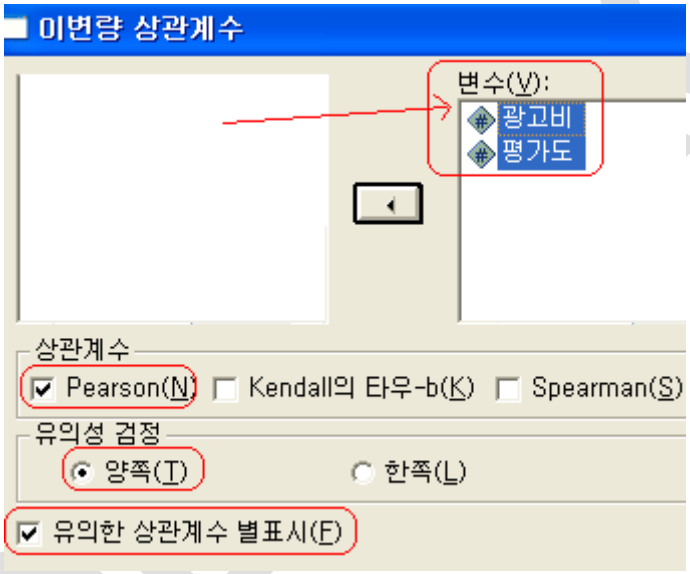
변수	N	평균	표준편차	합
rate	17	35.07647	28.07409	596.30000
spend	17	29.48824	22.95610	501.30000

관계식이 $\hat{\beta} = \sqrt{\frac{Syy}{Sxx}} r$ 이었으므로 $\sqrt{\frac{Syy}{Sxx}} r = \frac{28.07}{22.96} * 0.707 = 0.864$ (페이지 45)는 회귀계수의 추정치와 동일하다. 그러므로 단순회귀에서 상관계수의 유의성 검정($H_0 : r = 0$)과 회귀모형의 기울기 회귀계수 유의성 검정($H_0 : \beta = 0$)은 동일하다.

SPSS



상관계수를 구하려는 변수를 변수 타원으로 표시한 부분은 default로 나타나므로 따로 설정할 필요는 없다.



HOMEWORK #3 DUE 3월 23일(수)

CANCER.txt (텍스트 데이터) **SPSS와 SAS 모두 사용하여 분석하시오.**

연 평균 온도(F: Fahrenheit, 설명변수)가 여성 종양 사망지수(mortality index)에 영향을 미치는지 알아보기 위하여 유럽 몇 지역을 대상으로 조사한 자료이다. (SPSS 이용하기)

①산점도를 그리자. 산점도를 이용하여 이상치가 있으면 제외하자. 하나는 있다. 이를 제외한 다음 분석을 실시하자.

②회귀모형을 추정하고 유의성을 검정하시오.

③예측치와 잔차를 구하자.

④연 평균 온도가 90도 일 때 여성 종양 사망지수의 평균 예측치를 구하시오. 신뢰수준은 90%로 하시오.

⑤상관계수를 구하고 해석하시오. 회귀계수와의 관계를 밝히시오.