Chapter 15

Regression Analysis I

경영대학 재무금융학과

윤선중

Objectives

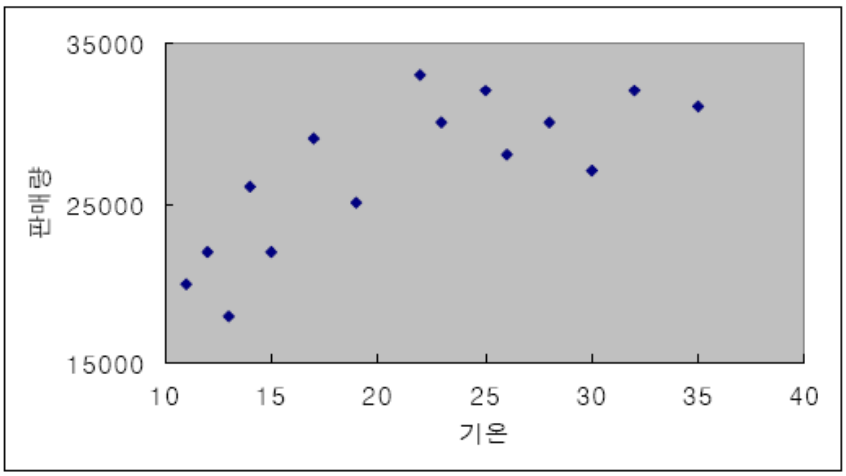
- 단순회귀분석(linear regression); 상관분석(correlation)
  - 두 구간변수간의 관계를 분석하기 위한 통계 기법
  - 회귀분석:
    - 선형관계
    - 최소자승법(least squares method)을 통해 절편 (intercept)과 기울기 (slope) 추정
  - 추정치의 표준오차 (standard error of estimate)
    - 선형관계의 충분한 증거를 확인하기 위해 기울기에 대한 검정이 이루어짐
  - 결정계수 (coefficient of determination)
    - 선형관계의 강도
  - 피어슨 상관계수
    - 정규분포를 따르는 두 변수간의 관계를 측정 및 검정

Introduction

- 정의
  - 서로 다른 변수들간의 선형관계를 나타내기 위하여 사용되는 분석기법
  - 특정 변수들의 값에 기초하여 다른 한 변수의 값을 추정하기 위하여 활용
- 종속변수 (dependent variable) vs. 독립변수 (independent variable)
  - _____: 다른 변수들의 값에 근거하여 추정되는 변수
  - _____: _____를 설명하기 위하여 사용되는 변수
- 회귀분석 vs. 상관분석
  - 회귀분석은 변수들 간의 인과관계에 초점
  - 상관분석은 변수들간의 단순한 상관관계에 초점

Example

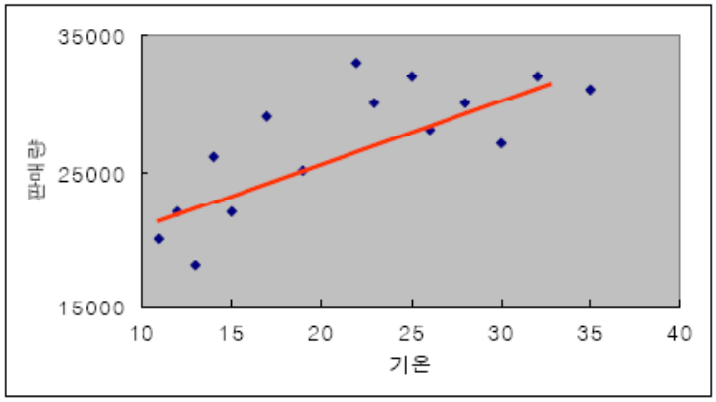
- 아이스크림 판매량
  - 기온과 아이스크림 판매량은 인과관계를 가지고 있을 거라고 예상
  - 조사한 결과 다음과 같은 그래프를 얻었음



  - 기온과 아이스크림 판매량은 어떠한 관계를 가지고 있는가?

Example

- 상관분석
  - 상관계수 = 0.762
- 회귀분석
  - 아래의 관계를 가장 잘 설명할 수 있는 함수식은 무엇인지 결정



  - Ex) 판매량 = 17,000+500*기온 + 오차항 ; Excel Graph 추세선 옵션으로 가능

Regression Model

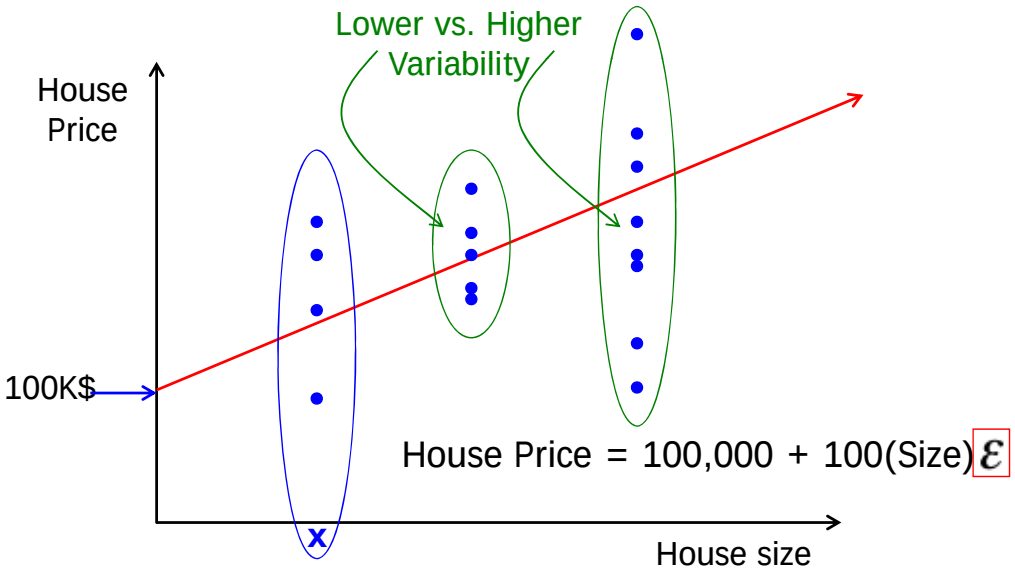
- 확률적 모형 (probabilistic model)
  - 독립변수의 값에 의해서 종속변수가 설명되지만, 임의성(randomness)을 포함하고 있는 모형
  - 즉, 종속변수를 (설명변수 + 오차항)에 의해서 설명하는 모형
  - 확정적 모형(deterministic model)의 반대개념
- 예제: 집을 짓는 비용
  - 확정적 모형:

    y = $100,000 + (100$/ft2)(x) ; x는 집의 크기
  - 확률적 모형:

    y = $100,000 + (100$/ft2)(x) + e ;
    - e는 오차항으로서 임의성 (randomness)을 대변함

Regression Model

- 확률적 모형의 예 – cont'd



Regression Model

- 선형회귀모형
  - 설명변수와 종속변수의 관계를 선형으로 나타낸 경우
  - 단순 선형회귀모형 (simple linear regression model)
    - 설명 변수가 한 개인 선형회귀모형
  - 다중 선형회귀모형 (multiple linear regression model)
    - 설명변수가 두 개 이상인 선형회귀모형
  - 비선형회귀모형의 반대개념
    - 설명변수와 종속변수의 관계를 비선형으로 나타낸 경우
    - 이차회귀모형, 로지스특(logistic) 회기모형 등

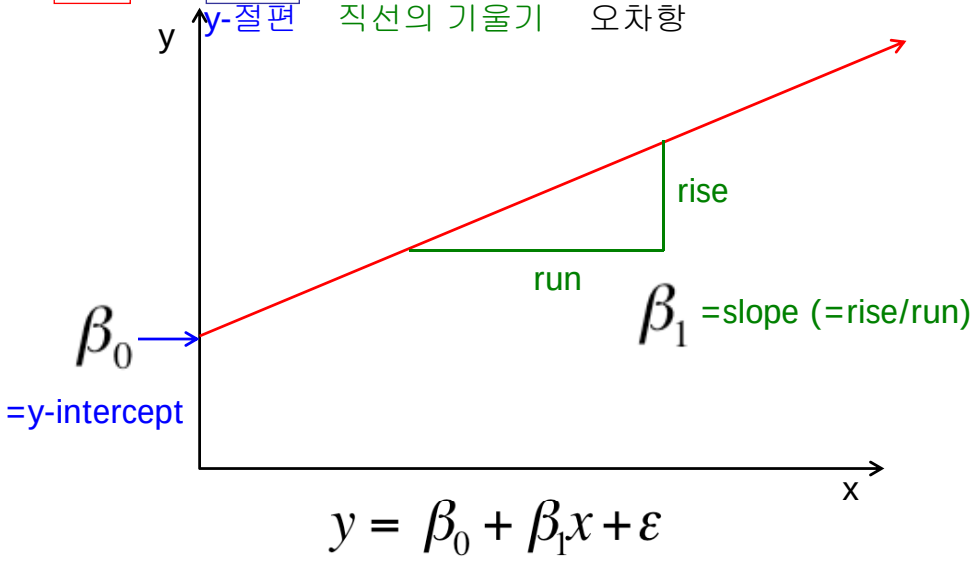
Regression Model

- 단순 선형회귀모형의 형태

종속변수 독립변수

$$y = \beta_0 + \beta_1 x + \varepsilon$$

y-절편 직선의 기울기 오차항



$$y = \beta_0 + \beta_1 x + \varepsilon$$

Regression Model

- 기타 회귀 모형

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2^{\,2} + \varepsilon$$

$$\log Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Estimation

- 추정의 의미
  - 회기 모형에서 회기 계수들을 추정

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon \quad \Rightarrow \quad Y = \hat{\beta}_0 + \hat{\beta}_1 X_1$$

- **최소자승법 (Least Square Method)**
  - 실제관측치와 회귀선간의 거리가 가장 가깝도록 추정하는 기법
  - 잔차의 제곱의 합을 최소화 하도록 추정
    - 잔차 (residual) = 실제관측치 – 회기선

Estimation

- **최소자승선 (Least Squares Line)**
  - 최소자승법을 이용하여 계산된 회기선

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$$

- 회귀계수의 결정

$$\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{\sum x_i y_i}{\sum x_i^2}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$x_i = X_i - \bar{X}, \quad y_i = Y_i - \bar{Y},$$

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{x^2}},$$

$$\text{단, } s_{xy} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}, s_{x^2} = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

Exercise 1

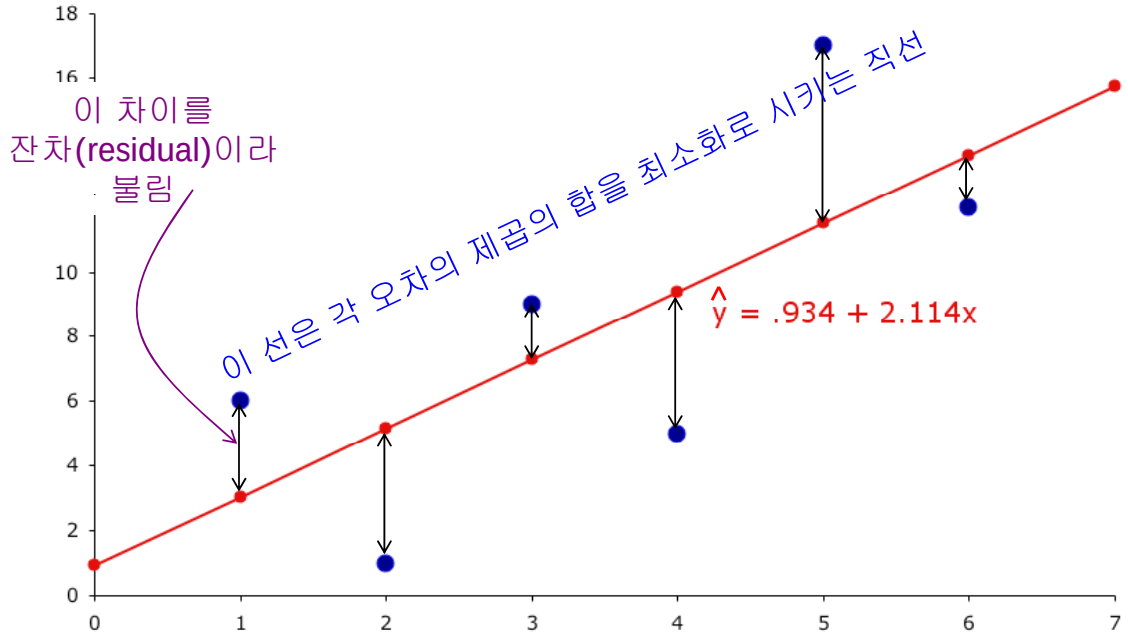
- 예제 1
  - 임의로 선택된 6명의 종업원의 근속년수와 연간 보너스

근속연수 x	1	2	3	4	5	6
보너스 ($1000) y	6	1	9	5	17	12

  - (1) 단순 선형회귀모형을 설정하시오
  - (2) 최소자승선을 구하시오
  - (3) 만약 근속년수가 4.5년인 종업원이 있다면, 연간 보너스는 얼마일 것으로 예측할 수 있는가?
  - (4) 회귀분석이 잘 이루어졌다는 사실은 어떻게 검증할 것인가?

Exercise 1

Example 15.1



Exercise 1

- 추정 방법
  - 1. 직접 계산: 회귀 계수를 구하는 공식에 의한 해결
    - 책 참고
  - 2. 엑셀의 그래프의 산포도 –추세선그리기 – 추세선 식 추가 옵션을 이용
  - 3. 데이터 분석 –regression (회기분석)을 이용
    - 예제 15.2에서 참고

Exercise 2

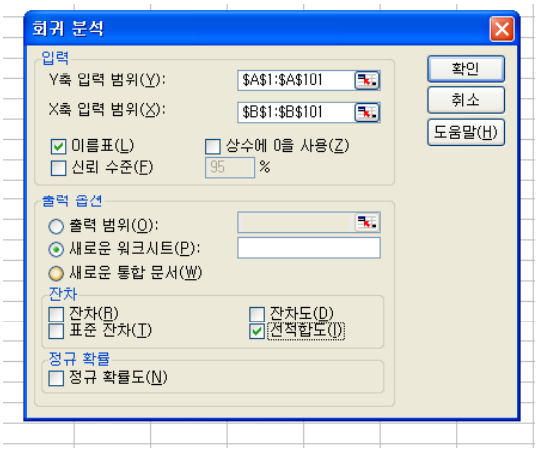
- 예제 15-02; Xm15-02: Toyota Camry 중고차의 주행거리와 가격, Part I
  - 방법 1: 직접 계산
  - 방법 2: Excel 이용

D10

	A	B	C
1	Price	Odometer	Color
2	14.6	37.4	1
3	14.1	44.8	1
4	14.0	45.8	3
5	15.6	30.9	3
6	15.6	31.7	2
7	14.7	34.0	2
8	14.5	45.9	1
9	15.7	19.1	3
10	15.1	40.1	1
11	14.8	40.2	1
12	15.2	32.4	2
13	14.7	43.5	1
14	15.6	32.7	1
15	15.6	34.5	2
16	14.6	37.7	2
17	14.6	41.4	1
18	15.7	24.5	3
19	15.0	35.8	1
20	14.7	48.6	1
21	15.4	24.2	1

Exercise 2

- **Excel – 데이터분석**
  - **Xm15-02 읽기**
  - **도구-데이터분석-회기분석 클릭**
  - **Y축입력범위(A1:A101), X축 입력범위 (B1:B101)입력**
  - **산포도를 그리기 위해서는 확인을클릭하기 전에 선적합도(line fit plot)을 클릭**



Exercise 2

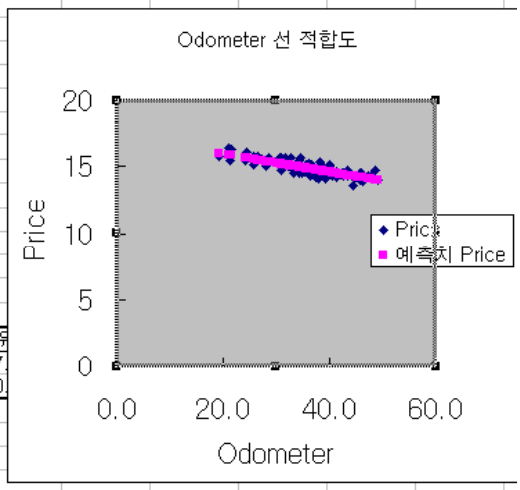
요약 출력

회귀분석 통계량	
다중 상관계수	0.805167979
결정계수	0.648295475
조정된 결정계수	0.644706653
표준 오차	0.326400626
관측수	100

분산 분석

	자유도	제곱합	제곱 평균	F 비	유의한 F
회귀	1	19.25560737	19.25561	180.643	5.75E-24
잔차	98	10.44629263	0.106595		
계	99	29.7019			

	계수	표준 오차	t 통계량	P-값	하위 95%	상위
Y 절편	17.24872734	0.182092574	94.72505	3.57E-98	16.88737	17.
Odometer	-0.066860885	0.004974639	-13.4403	5.75E-24	-0.07673	-0.



잔차 출력

관측수	예측치 Price	잔차
1	14.74813023	-0.14813023
2	14.25335968	-0.153359679
3	14.18649879	-0.186498793
4	15.18272599	0.417274015
5	15.12923728	0.470762723
6	14.97545724	-0.27545724
7	14.1798127	0.320187295
8	15.97168443	-0.271684432
9	14.56760584	0.53239416

Test on the Regression Coefficients

- 회귀계수에 관한 가설검정
  - 최소자승법이 최적의 직선을 제공하기는 하지만, 두 변수간의 관계가 존재하지 않거나, 비선형 관계가 존재할 수 있음.
  - 따라서 회기 계수에 대한 검증이 필요!!
  - 가설검증에 필요한 통계량
    - 추정치의 표준오차(standard error of estimate)
    - 결정계수 (coefficient of determination)
  - 검정법
    - 기울기의 t-검정 (t-test of the slope)
  - 오차제곱합(sum of squares for error) SSE에 기초

Test on the Regression Coefficients

- 회귀계수에 관한 가설검정

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon \quad \Rightarrow \quad Y = \hat{\beta}_0 + \hat{\beta}_1 X_1$$

  - Step 1: 가설설정: $H_0 : \beta_1 = 0$

    $H_1 : \beta_1 \neq 0$
  - Step 2: 표본통계량 계산: $\hat{\beta}_1$ 추정
  - Step 3: 통계적 의사결정

$$p-value = 2 \times P(\hat{\beta}_1 > a) = 2 \times P\left(\frac{\hat{\beta}_1 - 0}{s_{\hat{\beta}_1}} > \frac{a-0}{s_{\hat{\beta}_1}} \right) = 2 \times P\left(t > \frac{a-0}{s_{\hat{\beta}_1}} \right)$$

    - 단, t분포의 자유도는 n-2

R-square (R2)

- 오차제곱합 (sum of squares for errors)와 추정치의 표준오차

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \qquad s_\varepsilon = \sqrt{\frac{SSE}{n-2}}$$

  - SSE로써 회귀모형의 설명력을 평가할 수 있는가?

- 결정계수 (R^2)

$$R^2 = 1 - \frac{SSE}{TSS} = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y}_i)^2}, \qquad \therefore 0 \le R^2 \le 1$$

  - 종속변수의 총 변동 중에서 회귀모형에 의하여 설명된 변동의 비율
  - R^2를 이용하여 회귀모형의 설명력을 평가하는 것이 일반적!

Exercise 2

- Xm 15-02; 예제 15-2; 중고차 가격과 주행거리 Part II: 기울기 계수 검정
  - 회귀분석실행

회귀분석 통계량	
다중 상관계수	0.805167979
결정계수	0.648295475
조정된 결정계수	0.644706653
표준 오차	0.326488626
관측수	100

분산 분석

	자유도	제곱합	제곱 평균	F 비	유의한 F
회귀	1	19.25560737	19.25561	180.643	5.75E-24
잔차	98	10.44629263	0.106595		
계	99	29.7019			

	계수	표준 오차	t 통계량	P-값	하위 95%	상위 95%	하위 95.0%	상위 95.0%
Y 절편	17.24872734	0.182092574	94.72505	3.57E-98	16.88737	17.61008	16.88737	17.61008
Odometer	-0.066860885	0.004974639	-13.44035	5.75E-24	-0.076733	-0.056989	-0.076733	-0.056989

Exercise 2

- 중고차의 주행거리와 가격은 선형 관계를 가지고 있는가?: 기울기 계수에 대한 검정
  - **예제 15.4 & 15.5**
    - H1: $\beta 1 \neq 0$, H0: $\beta 1 = 0$

$$t = \frac{b_1 - \beta_1}{s_{b_1}} \qquad s_{b_1} = \frac{s_\varepsilon}{\sqrt{(n-1)s_x^2}}$$

    - 5% 신뢰도에서

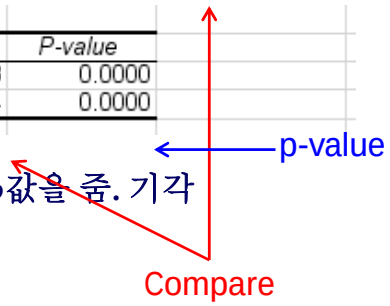
$$s_{b_1} = \frac{s_e}{\sqrt{(n-1)s_x^{\,2}}} = \frac{.3265}{\sqrt{(99)(43.509)}} = .00497$$

$$t = \frac{b_1 - \beta}{s_{b_1}} = -13.46$$

$$t < -t_{\alpha/2,\nu} = -t_{.025,98} \approx -1.984 \quad or \quad t > t_{\alpha/2,\nu} = t_{.025,98} \approx 1.984$$

Exercise 2

		Coefficients	Standard Error	t Stat	P-value
17	Intercept	17.25	0.182	94.73	0.0000
18	Odometer	-0.0669	0.0050	-13.44	0.0000



- **–13.49는 -1.984보다 매우 작아서 0과 가까운 p값을 줌. 기각**

- 결정계수

$$R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2}, \; or \; R^2 = 1 - \frac{SSE}{\sum (y_i - \bar{y})^2}$$

  - **Variation in y = SSE + SSR**

Exercise 2

$$s_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right] = \frac{1}{100-1}\left[53{,}155.9 - \frac{(3{,}601.1)(1{,}484.1}{100}\right] = -2.909$$

$$R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{(-2.909)^2}{(43.509)(.3000)} = .6483$$

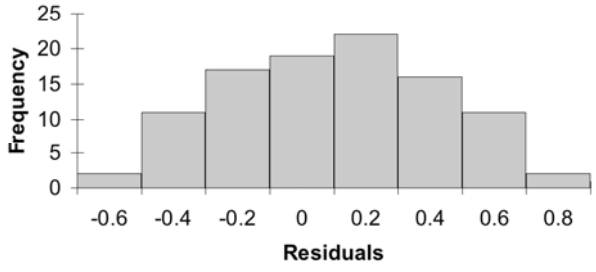
$$s_x^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right] = \frac{1}{100-1}\left[133{,}986.59 - \frac{(3{,}601.1)^2}{100}\right] = 43.509$$

$$s_y^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right] = \frac{1}{100-1}\left[22{,}055.23 - \frac{(1{,}484.1)^2}{100}\right] = .3000$$

회귀분석 통계량	
다중 상관계수	0.805167979
결정계수	0.648295475
조정된 결정계수	0.644706653
표준 오차	0.326488626

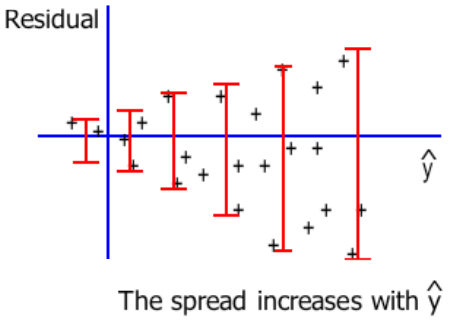
Residual Analysis

- 오차항의 필요조건
  - 만약 회귀모형이 제대로 설계되었다면,
    - 확률분포가 정규분포를 따라야 함
    - 기대값은 0이며, 표준편차는 항상 일정한 값이어야 함
    - 오차항 간에 자기 상관은 존재하지 않음
- 오차항의 필요조건 검증
  - 잔차(residual)를 이용하여 오차항의 필요조건을 검증
  - (1) 정규분포 조건

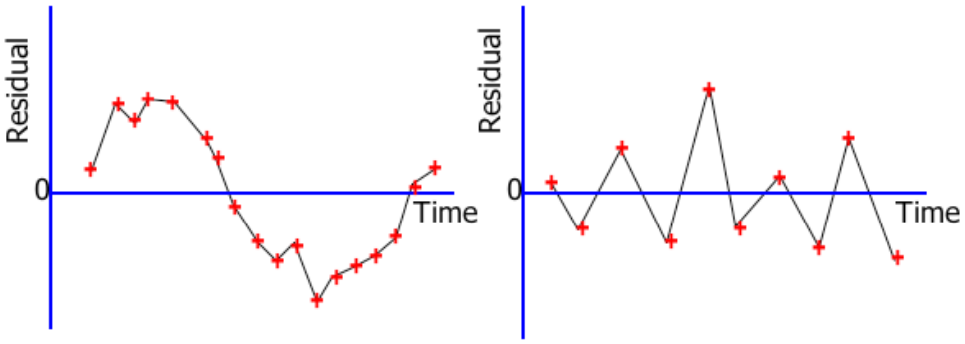


Residual Analysis

- (2) 동분산 조건

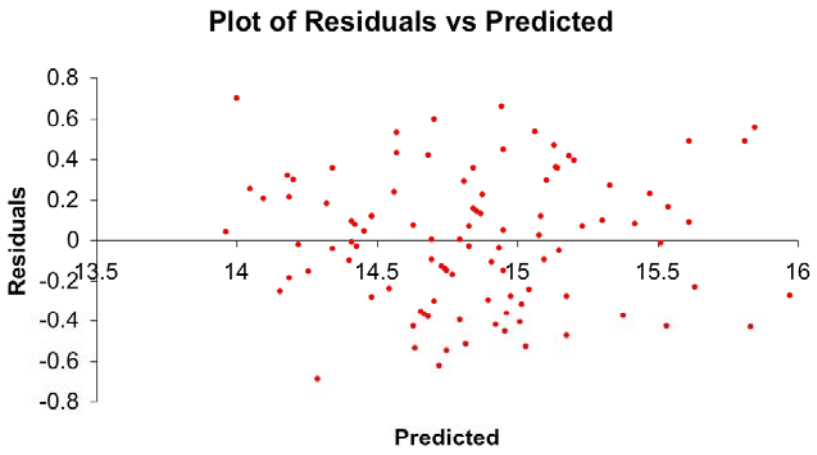


- (3) 비 자기상관(No auto-correlation)



Residual Analysis

■ 이상적인 오차상



Exercise 3

- 연습문제 15-49
  - 코카콜라 주가와 S&P500 지수간의 관계
  - 시장모형을 추정하시오

$$R_i = \beta_{i0} + \beta_{i1} X_M + \varepsilon$$

  - 회귀모형은 제대로 설계되었는지 다음의 각도에서 분석하시오
    - (1) 회귀계수에 대한 가설검정
    - (2) 결정계수 분석
    - (3) 잔차 분석

Discussion

- 회귀분석의 실행 순서
  - (1) 인과관계에 있다고 판단되는 변수를 설정
  - (2) 회귀모형을 구성
  - (3) 데이터를 수집
  - (4) 산점도를 그려 극단치(outlier) 제거
  - (5) 회귀모형 추정
  - (6) 회귀모형 적합도 평가
    - 회귀계수의 유의성/ 잔차분석/ 결정계수
  - (7) 회귀모형에 근거한 예측