제10장 단순회귀분석

회귀분석 (Regression analysis)이란?

- 독립변수(independent variable, 설명변수)와 종속변수(dependent variable, 반응변수)간의 관계를 검증하여 독립변수가 종속 변수에 미치는 영향력을 회귀식을 통해 알아보는 통계분석 방법

- 회귀분석의 두 가지 종류

1. 단순회귀분석 : 독립변수 1개와 종속변수 1개

    e.g) 한낮 최고 기온 (독립변수)과 아이스크림 매출액 (종속변수)과의 관계 분석

    아이스크림 매출액 = 13000+9000*한낮 최고 기온

2. (다)중회귀분석 : 독립변수 2개 이상과 종속변수 1개

    e.g) 일일 운동시간, 칼로리 섭취량 (이상 종속변수) 과 몸무게와의 관계분석 (독립변수)

    몸무게 = 60 – 3*일일운동시간 + 2*칼로리섭취량

※ 종속변수가 범주형인 경우 로지스틱 회귀분석을 사용

회귀분석의 목적

1. 독립변수와 종속변수간의 관계의 강도와 방향을 나타내는 회귀계수에 대한 추정 및 검정
2. 독립변수와 종속변수간의 관계를 설명하는 회귀방정식 모형을 찾음
3. 특정 독립변수가 다른 독립변수에 비해 얼마나 종속변수에 영향을 미치는지 파악
4. 독립변수로 종속변수를 예측

기호의 정의

$$Y = \alpha + \sum_{i=1}^{n} \beta_i X_i \quad \cdots \quad \text{회귀방정식(모집단)}$$

$$\hat{Y} = a + \sum_{i=1}^{n} b_i X_i \quad \cdots \quad \text{추정회귀방정식(표본)}$$

- X_i : 독립변수
- Y : 종속변수

단순회귀의 예

- 일일 최고기온과 냉면 판매량

날짜	(X)	(Y)
1일	29	77
2일	28	62
3일	34	93
4일	31	84
5일	25	59
6일	29	64
7일	32	80
8일	31	75
9일	24	58
10일	33	91
11일	25	51
12일	31	73
13일	26	65
14일	30	84

- 두 변수간의 관계를 예측하는 방법
  - 산점도 그리기
  - 상관관계 분석 하기

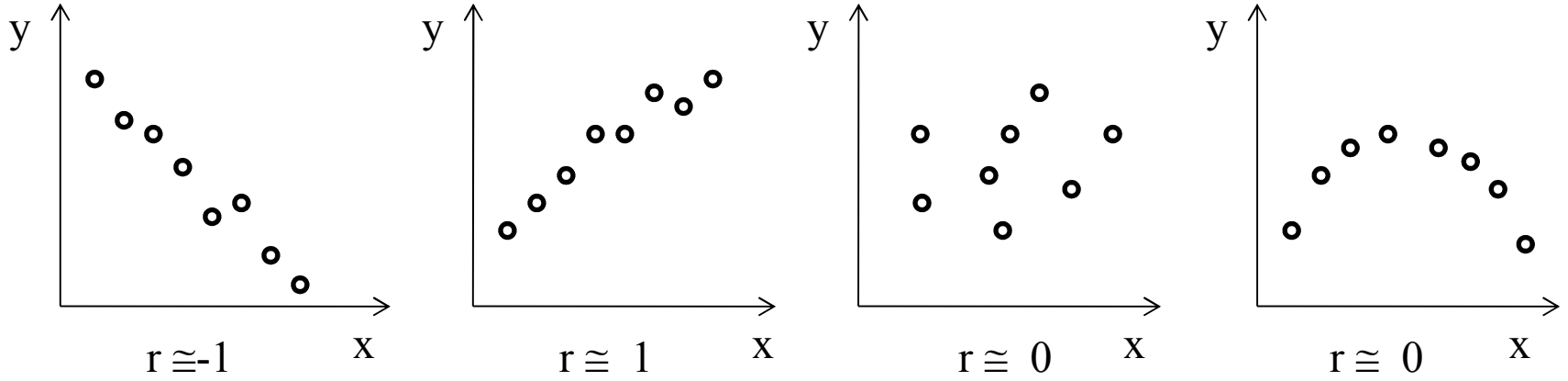
단순회귀 분석과 활용의 순서

1. 종속변수와 독립변수를 구분
2. 회귀분석에 기본 가정이 맞는지 분석
3. 독립변수와 종속변수간의 산점도 (scatter plot)을 작성
4. 상관관계 분석
5. 회귀식을 구함
6. 회귀식의 적합도 검정(F – 검정)
7. 회귀계수의 유의성 검정 시행 (β=0 인지 검정)
8. 유의한 결과가 나오면 모회귀식의 추정
9. 예측에 사용

- 상관관계분석과 회귀분석의 차이점
  - 상관관계분석 : 단순히 두 변수 사이의 상관 관계 정도만 분석
  - 회귀 분석 : 두 변수 사이의 인과 관계를 알 수 있고, 이를 통해 한 변수로부터 다른 변수의 변화 정도를 예측할 수 있는 분석 방법

단순회귀 분석과 활용의 순서 – 산점도 그리기

- 산점도 : 독립변수와 종속변수를 2차원상의 점으로 표현한 그림



- 산점도에서 주요 관찰 사항
  1. 변수 사이에 관찰되는 관계는 어떠한가 (직선, 곡선, 관계없음)
  2. 이상치 (Outlier)는 없는가
  3. 몇 개의 그룹으로 나뉠 수 있는가?

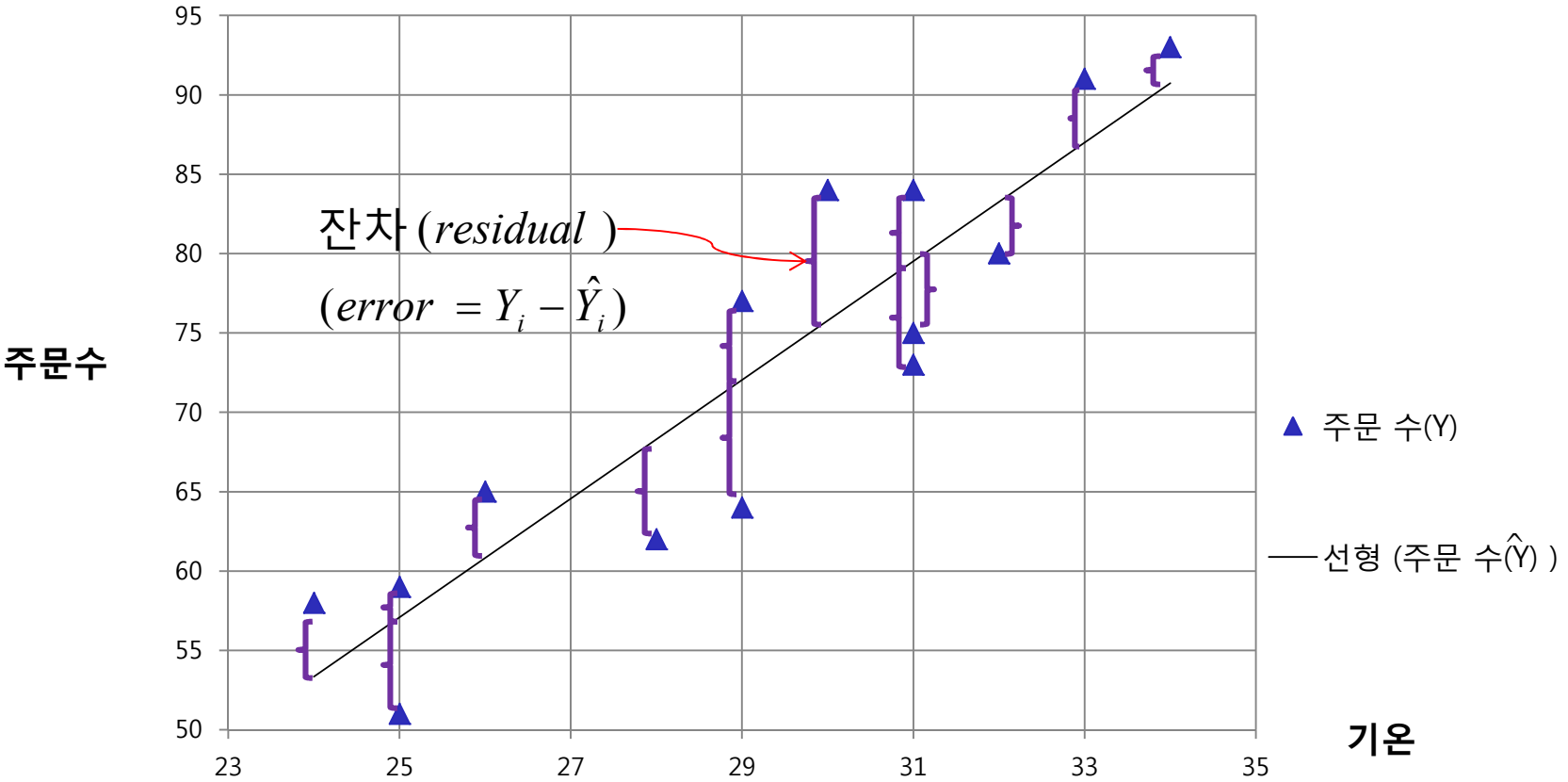
단순회귀 분석과 활용의 순서 – 상관관계 분석

- 이전 장 (상관관계 부분 참고)
- 상관관계는 직선의 상관관계 여부를 상관계수(r) 로 표시
  - $-1 \leq r \leq 1$
  - r=-1 은 음의 상관관계
  - r=0 는 상관관계 없음
  - r=1 은 양의 상관관계

- SPSS 분석 시 Pearson 상관계수와 유의 확률을 이용하여 상관관계 여부를 검정

단순회귀 분석과 활용의 순서 – 회귀식 구하기

- 회귀식의 원리
  - 최소제곱법 (최소자승법) : 잔차 제곱의 합을 최소로 하는 직선을 구함 , min e^2



단순회귀 분석과 활용의 순서 –모형 적합성 분석

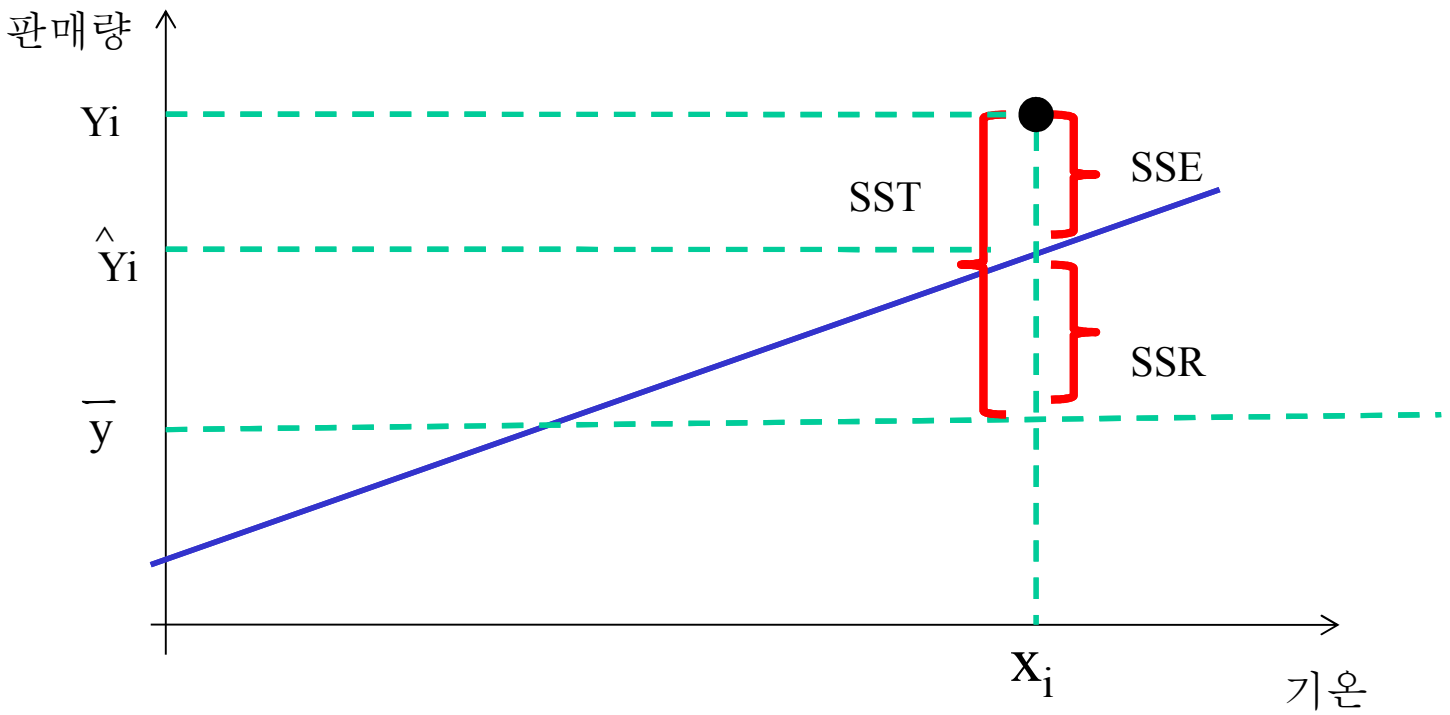
- 회귀 모형의 적합성 (모형이 의미가 있는지)을 검정
  - 독립변수가 종속변수를 설명하는데 도움이 되는지 검정
  - 분산분석을 이용 (분산분석 부분 참고)

요인	자유도(d)	제곱합 (SS)	평균제곱합 (Mean SS)	검정통계량 (F)	유의확률 (p-value)
회귀모형(Reg)	1	SSR	MSReg=SSR/k	F=MSR/MSE	F>F(1, n-2)
잔차 (error)	n-2	SSE	MSE=SSE/(n-k-1)		
합계 (Total)	n-1	SST			

- K는 독립변수의 수 (단순회귀에서 k=1)

$$\frac{SSR/k}{SST/(n-k-1)}$$

단순회귀 분석과 활용의 순서 -모형 적합성 분석



- SST=SSE+SSR
- SST : 개별 데이터와 전체 평균의 거리 (종속변수의 평균에 기초한 오차제곱합)
- SSE : 회귀식으로도 설명 못하는 부분 (회귀식에 기초한 오차제곱합)
- SSR : 회귀식으로 설명되는 부분

모형 적합성 분석의 이해

- 가정) 100명 학생 의 키와 몸무게를 측정하였음, 키와 몸무게는 선형의 관계가 있음, 몸무게는 = -90+0.9*키
- 한 명의 특정 학생을 뽑아 몸무게를 맞추어야 하는 게임을 함
  - 키를 모르는 경우 : 전체 100명의 학생의 평균 몸무게로 추정하는 것이 합리적임 (점 추정의 원리)
  - 키를 알 수 있는 경우 : 키를 이용하여 몸무게를 구하는 공식에 대입한 값으로 추정 (선형관계 이용)
- SST는 선형관계를 모르는 경우에 발생한 오차SSR는 직선으로 추정하여 발생한 오차 (회귀식으로 설명되는 부분)
- $\frac{SSR/k}{SST(n-k-1)}$ 는 회귀식(독립변수)을 이용하여 종속변수가 얼마큼 잘 예측되는 가를 나타냄

단순회귀 분석과 활용의 순서 – 회귀계수 유의성 분석

- 모 회귀계수 (기울기, β)가 0인지 검정 (중회귀에서는 다수의 회귀계수가 존재)
  - Ho : 기울기 (β)는 = 0 이다.
  - H1 : 기울기 (β)는 ≠ 0 이다.
- 기울기(β)가 0이면 독립변수가 종속변수의 예측에 영향을 미치지 못함 ⇩ 모형이 적합하지 않음
- 단순회귀에서는 모형적합성 검정과 동일
- 중회귀에서는 모형의 적합성 분석과 회귀계수 유의성 분석이 동일하지 않음

결정계수 (기여율)

- 모형이 적합하고 회귀계수 또한 유의한 경우 모형의 설명력을 알아볼 필요가 있음
- 결정계수 : 독립변수가 종속변수의 변동 중 얼마나 설명할 수 있는지를 나타냄
- R^2 로 표현
- R^2 = SSR/SST (=1-SSE/SST)로 계산

  e.g) R^2 = 0.85 는 전체 종속변수의 변동 중 85%는 독립변수로 설명이 된다는 의미

※ 중상관계수 (R)는 직선과 점이 얼마나 가까이 나타냄

- 측정값과 예측값의 상관계수로 계산
- 중상관계수의 제곱이 결정계수

회귀분석의 기본 가정

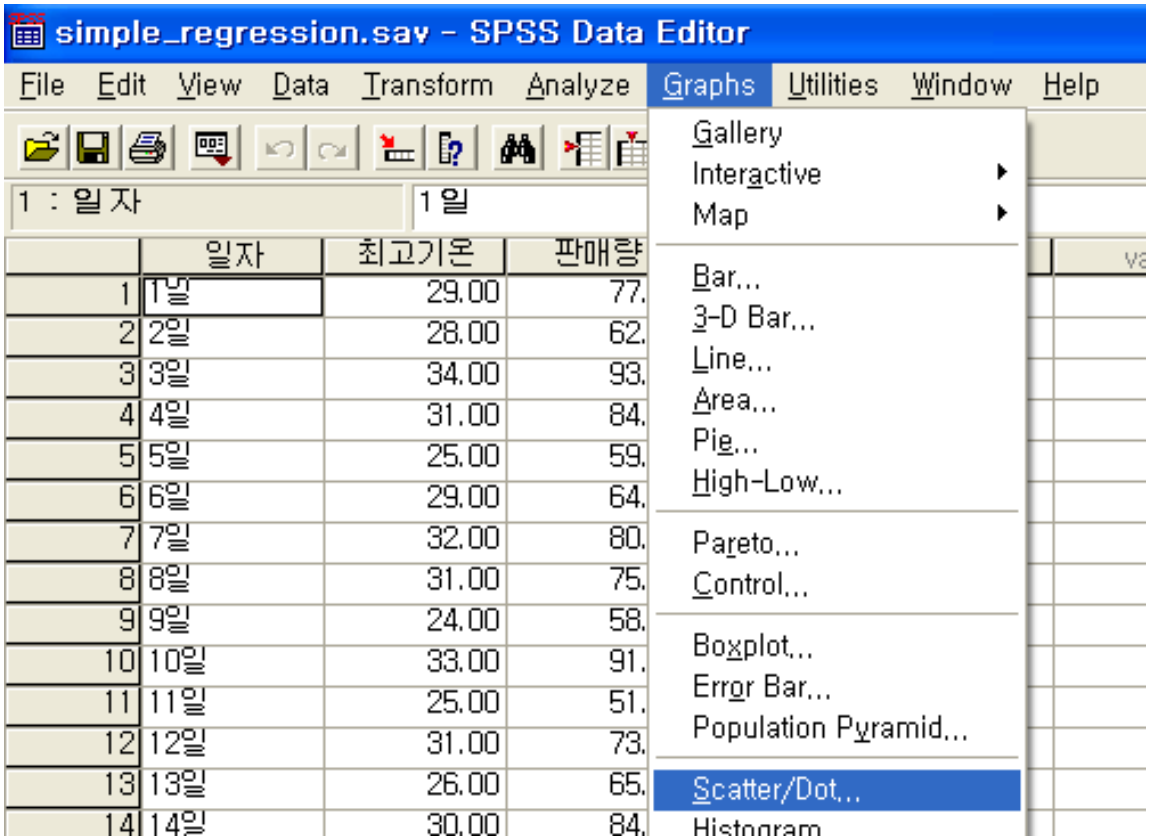
1. 선형성 : 두 변수간의 관계가 직선 관계
2. 정규성 : 같은 X 값을 가지는 Y는 정규분포를 따름
3. 독립성 : (X, Y) 의 각 데이터들과 오차들 간은 독립적
4. 등분산성 : 모든 X에서 Y의 분산은 모두 동일

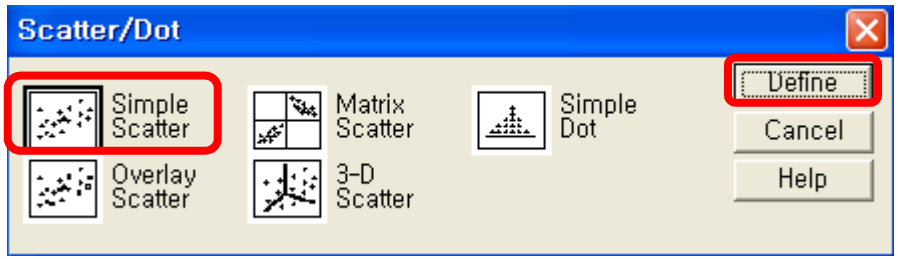
- 위의 가정의 맞는지 모두 밝히기 힘듦
  - 선형성이 위배되면 데이터를 변환하여 선형적으로 바꿈
  - 정규성은 충분한 데이터(표본 수) 확보
  - 독립성은 잔차 plot과 자기상관(계열상관) 통계량을 통해
  - 잔차분석을 이용하여 등분산성이 맞는지

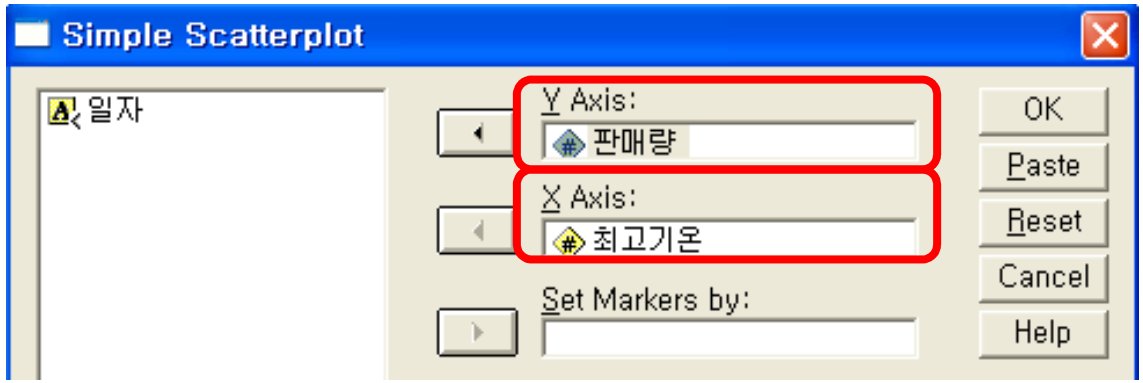
실습

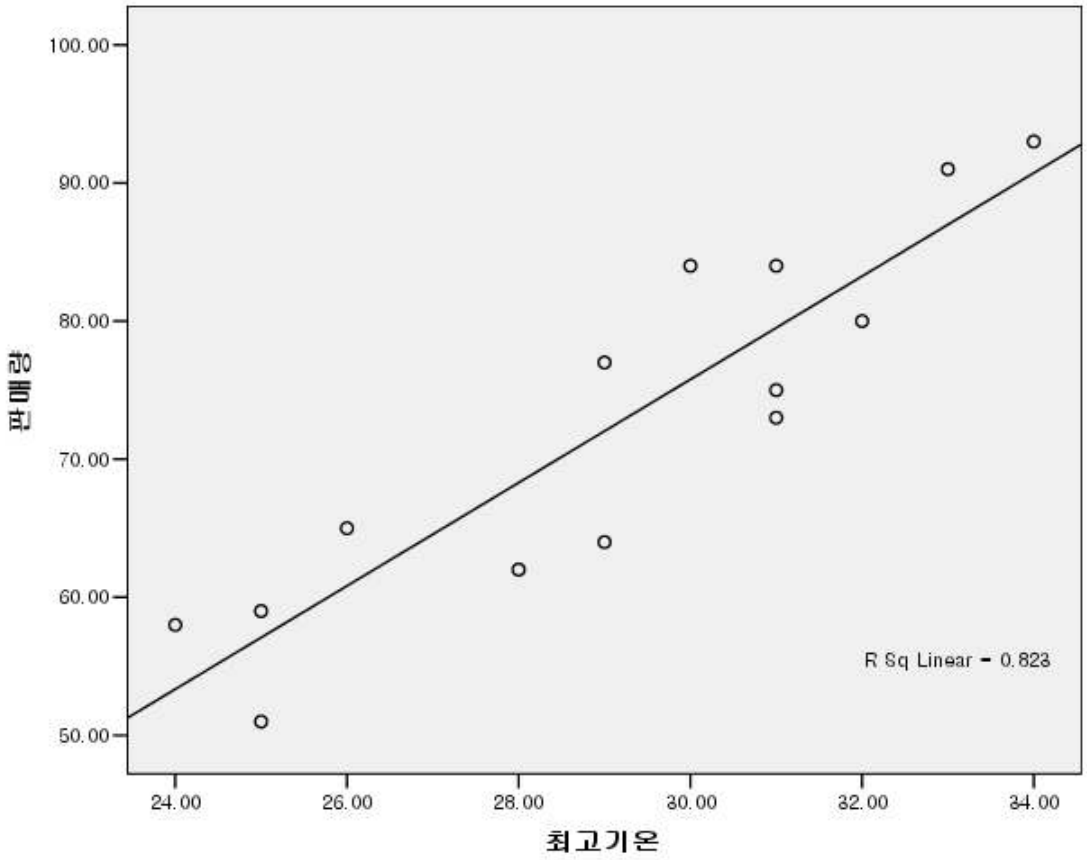
- SPSS를 통한 단순회귀분석 실습
  - 데이터 : simple_regression.sav
  - 실습 주요내용
    - 산점도 그리기를 통한 데이터의 직선관계 이해
    - 상관관계 분석을 통한 상관계수와 상관여부 이해
    - 회귀분석을 통한 회귀식 도출
    - 회귀식 적합성 검정
    - 잔차 분석
    - 예측

파일열기 및 산점도 그리기

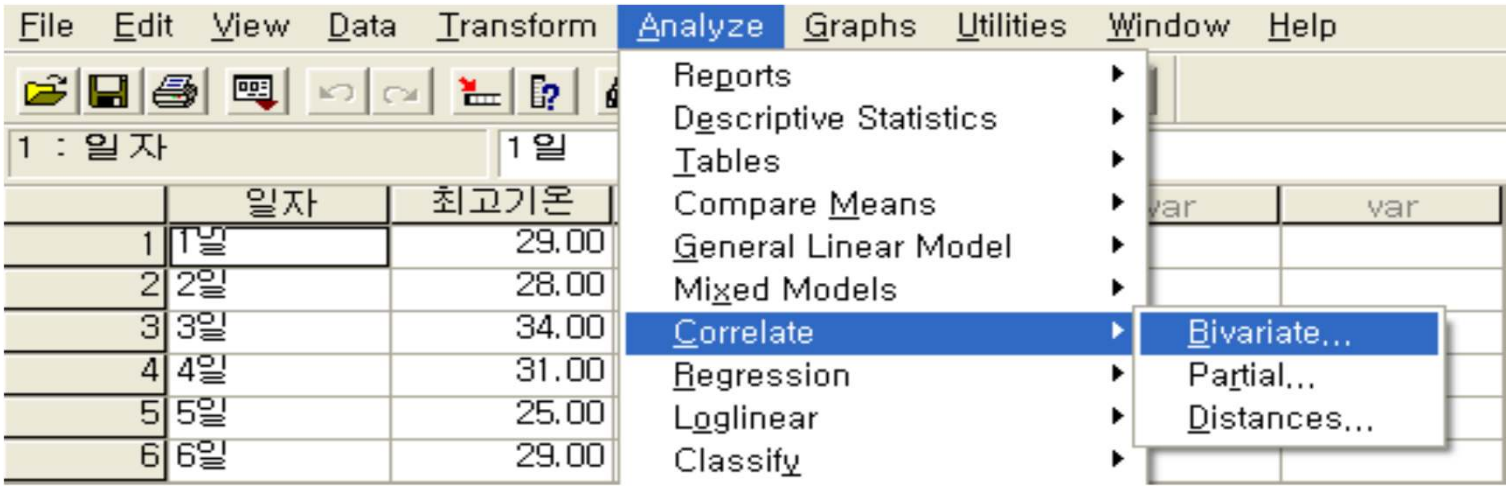





Simple Scatterplot
일자
Y Axis:
판매량
X Axis:
최고기온
Set Markers by:
OK
Paste
Reset
Cancel
Help


100.00
90.00
80.00
70.00
60.00
50.00
판매량
24.00
26.00
28.00
30.00
32.00
34.00
최고기온
R Sq Linear = 0.823

상관분석

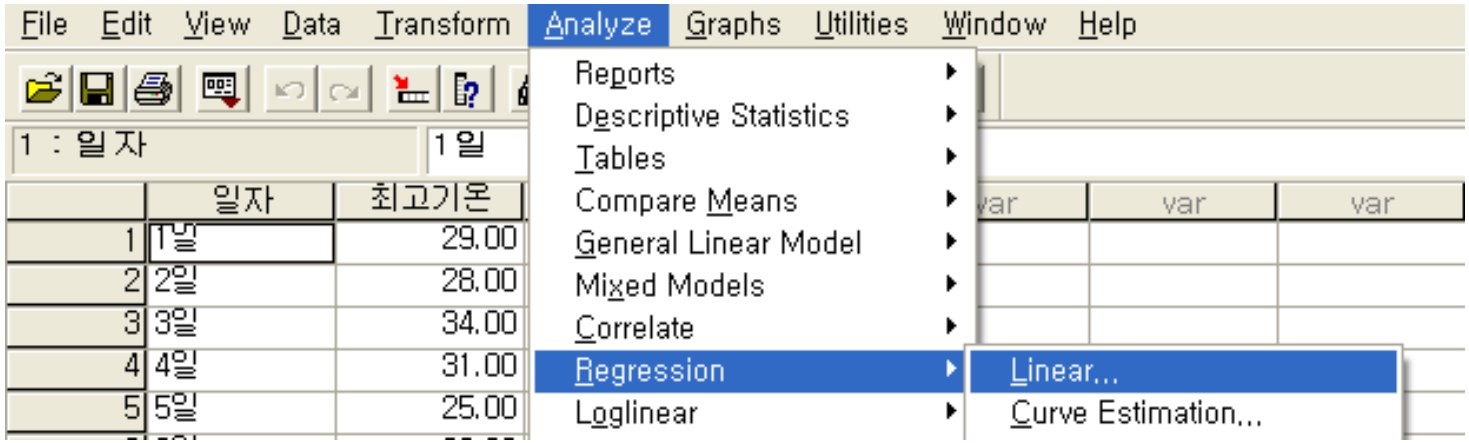


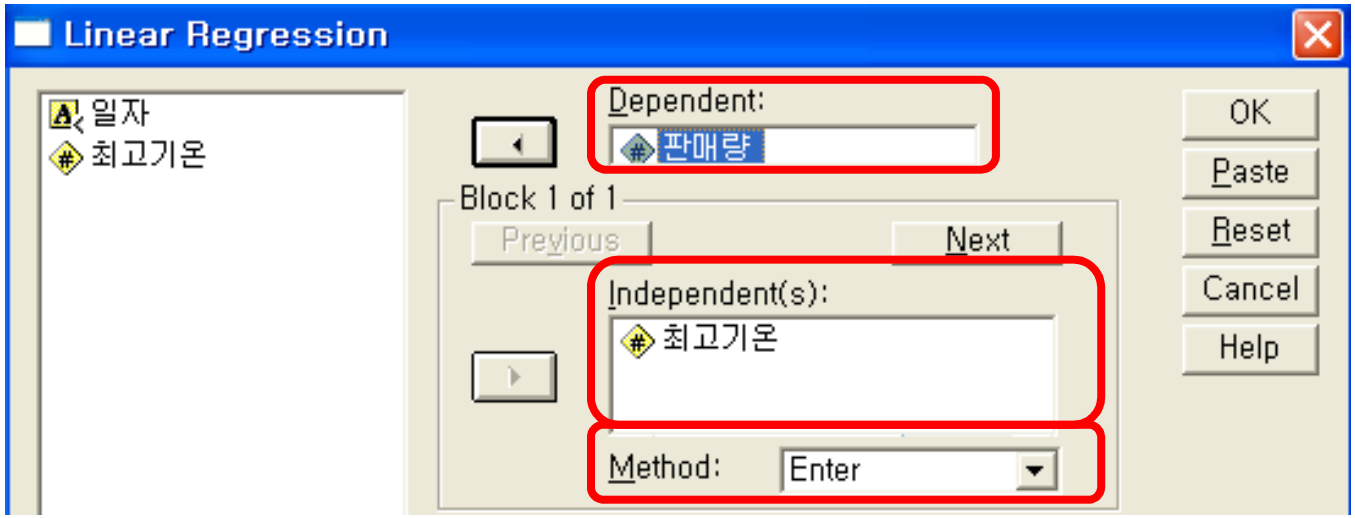
Correlations

		최고기온	판매량
최고기온	Pearson Correlation	1	.907**
	Sig. (2-tailed)		.000
	N	14	14
판매량	Pearson Correlation	.907**	1
	Sig. (2-tailed)	.000	
	N	14	14

**. Correlation is significant at the 0.01 level

단순선형회귀

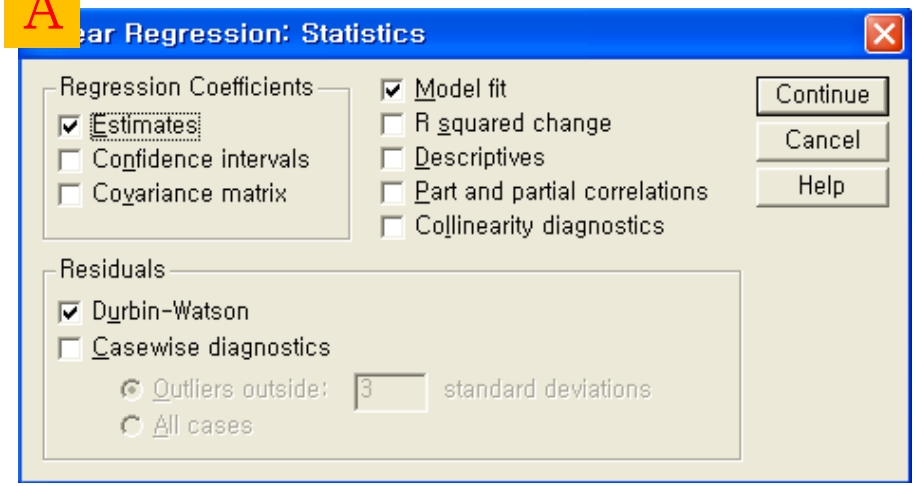

File Edit View Data Transform Analyze Graphs Utilities Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Linear...
Curve Estimation...
1 : 일자
1일
일자
최고기온
1 1일 29.00
2 2일 28.00
3 3일 34.00
4 4일 31.00
5 5일 25.00
var
var
var


Linear Regression
일자
최고기온
Dependent:
판매량
Block 1 of 1
Previous
Next
Independent(s):
최고기온
Method:
Enter
OK
Paste
Reset
Cancel
Help

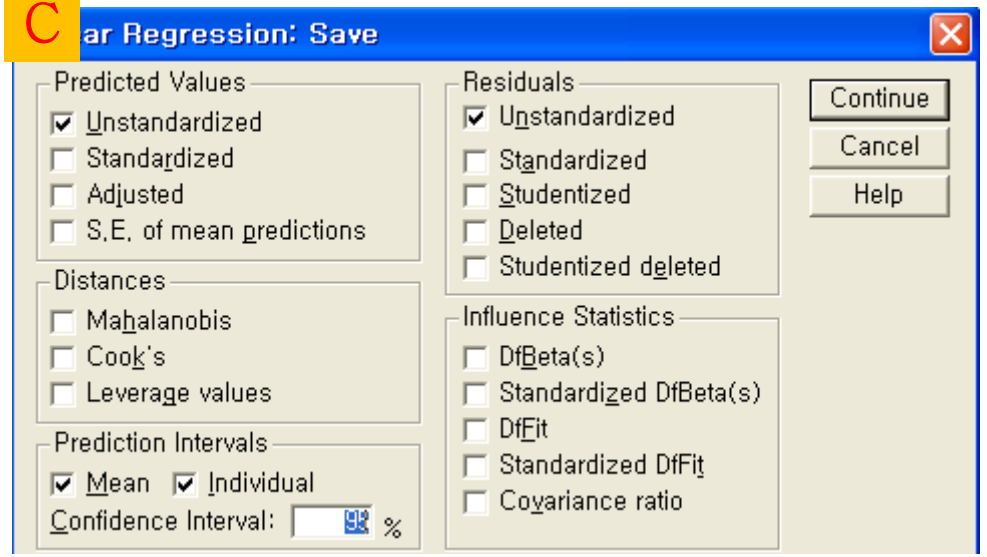

Statistics...
Plots...
Save...
Options...

A B C

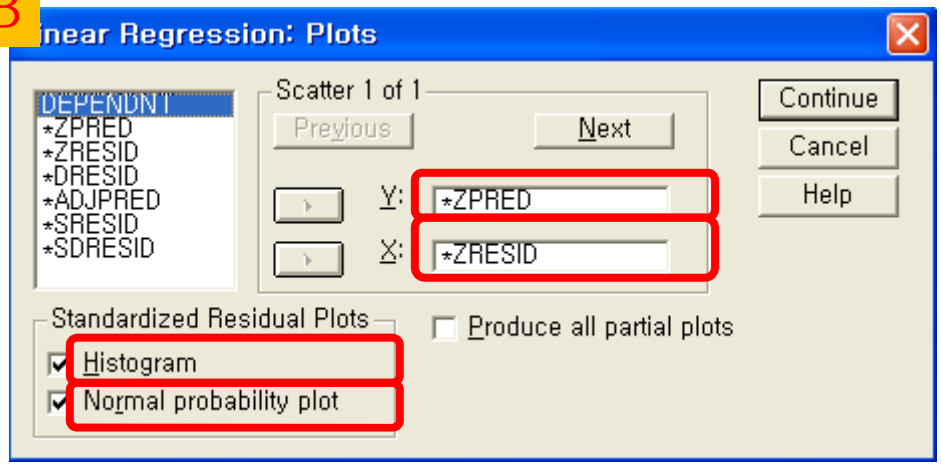
A


ear Regression: Statistics
Regression Coefficients
Estimates
Confidence intervals
Covariance matrix
Model fit
R squared change
Descriptives
Part and partial correlations
Collinearity diagnostics
Continue
Cancel
Help
Residuals
Durbin-Watson
Casewise diagnostics
Outliers outside:
3
standard deviations
All cases

C


ar Regression: Save
Predicted Values
Unstandardized
Standardized
Adjusted
S.E. of mean predictions
Residuals
Unstandardized
Standardized
Studentized
Deleted
Studentized deleted
Continue
Cancel
Help
Distances
Mahalanobis
Cook's
Leverage values
Influence Statistics
DfBeta(s)
Standardized DfBeta(s)
DfFit
Standardized DfFit
Covariance ratio
Prediction Intervals
Mean
Individual
Confidence Interval:
95
%

B


near Regression: Plots
DEPENDNT
*ZPRED
*ZRESID
*DRESID
*ADJPRED
*SRESID
*SDRESID
Scatter 1 of 1
Previous
Next
Y:
*ZPRED
X:
*ZRESID
Continue
Cancel
Help
Standardized Residual Plots
Histogram
Normal probability plot
Produce all partial plots

Variables Entered/Removed [b]

Model	Variables Entered	Variables Removed	Method
1	최고기온 [a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: 판매량

Model Summary [b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.907[a]	.823	.808	5.70882	1.669

a. Predictors: (Constant), 최고기온 모델의 설명력

b. Dependent Variable: 판매량

자기상관

$d_L = 1.08$, $d_U = 1.36$

ANOVA [b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1812.340	1	1812.340	55.609	.000[a]
	Residual	391.088	12	32.591		
	Total	2203.429	13			

a. Predictors: (Constant), 최고기온

b. Dependent Variable: 판매량

모델의 적합성

Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-36.361	14.687		-2.476	.029
	최고기온	3.738	.501	.907	7.457	.000

a. Dependent Variable: 판매량

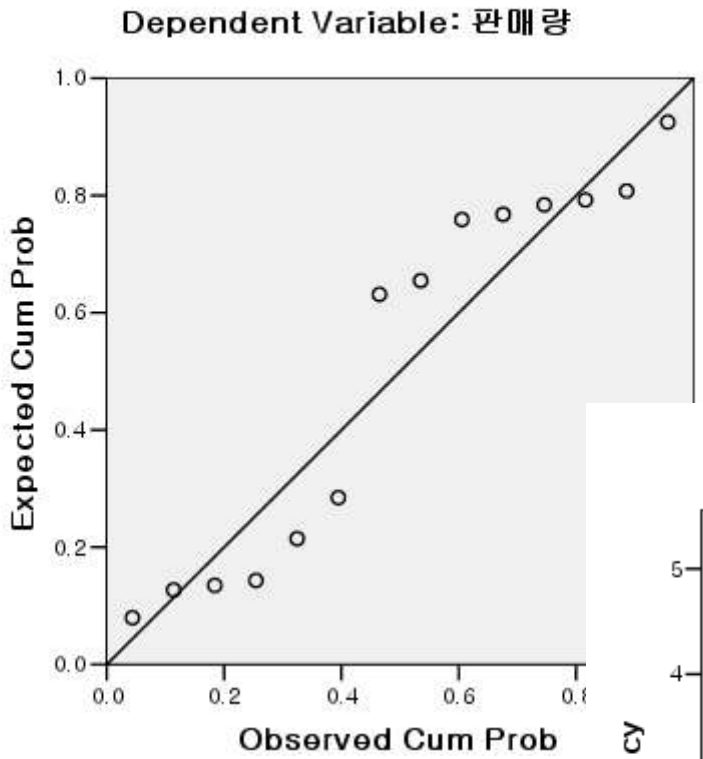
절편과 회귀계수의 유의성

- 회귀식은 적합함

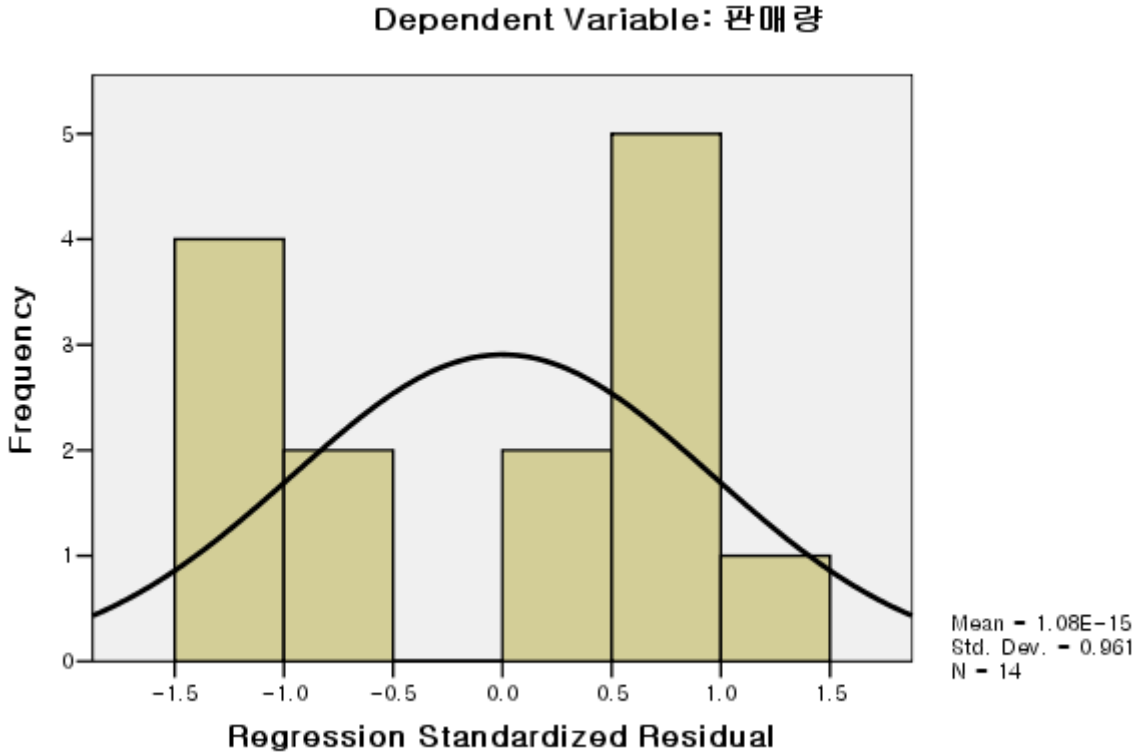
판매량 = -36.361+3.738*최고기온

- 기온이 35가 되었을 때 예측판매량은 ?
  판매량 = -36.361+3.738*35 = 94.5

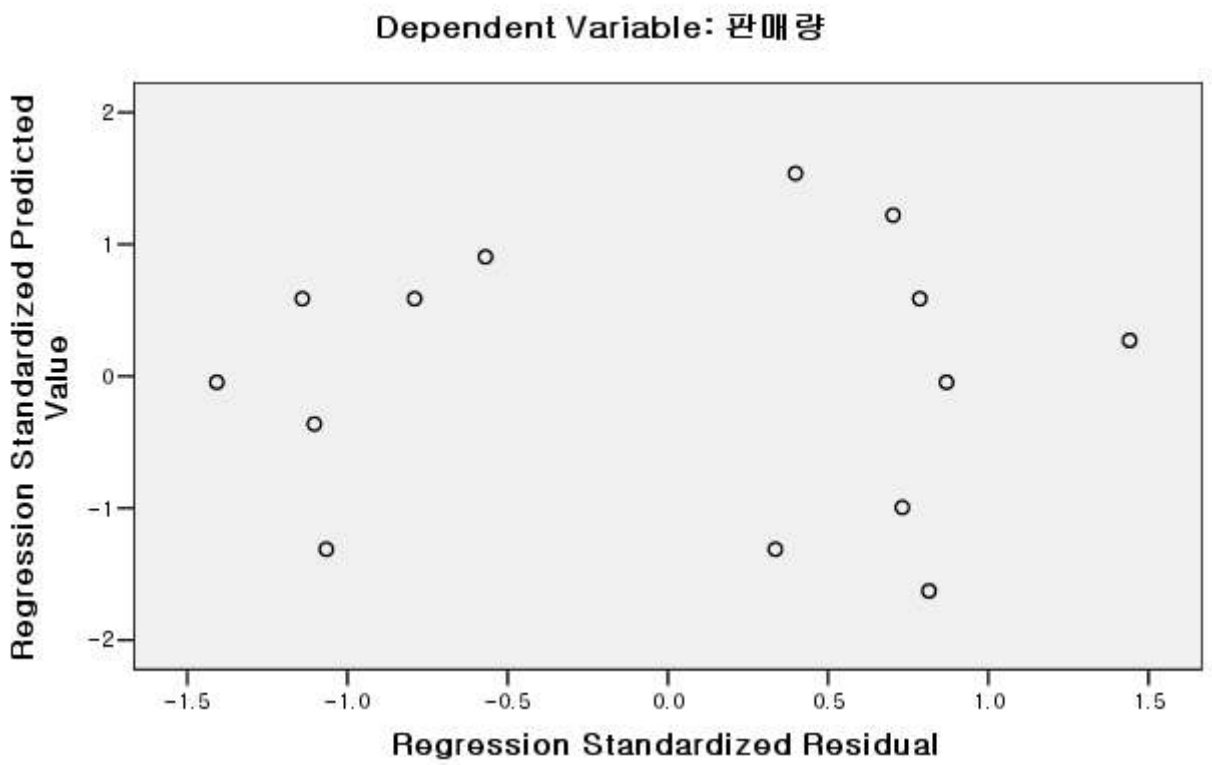
Normal P-P Plot of Regression Standardized Residual



- 자기상관이 의심됨 : 양의 자기상관
- Dubin-Watson's d 통계량에서 자기상관없음을 확인



- 관측 값의 개수가 작아서 잔차의 정규분포를 확인하기 힘듦



•표준화된 예측치에 따른 표준화된 잔차의 분포

➊ 잔차의 정규성, 독립성, 등분산성을 시각적으로 확인

연습문제

- 다음의 데이터를 이용하여 단순회귀 분석을 하시오.
  - 데이터 : simple_regression_ex.sav
  - 영업사원의 영업관련 적성검사 성적과 실제 판매실적 데이터
  - 분석 내용
    1. 산점도와 상관분석 후에 선형의 상관관계가 있는지 확인
    2. 회귀식의 추정 및 회귀식의 적합성 확인
    3. 회귀계수의 95% 신뢰구간 확인
    4. 자기회귀 여부 확인(d_L = 1.20, d_U = 1.41)
    5. P-P plot 작성
    6. 표준화된 예측치에 따른 표준화된 잔차의 분포도를 그리고 잔차의 정규성, 독립성, 등분산성을 시각적으로 설명
    7. 시험성적이 50인 경우 판매실적 예측