제11장 다중회귀분석

중회귀분석 (Regression analysis)이란?

- **2개 이상의 독립변수(independent variable, 설명변수)와 1개의 종속변수 (dependent variable, 반응변수)간의 관계를 검증하여 독립변수가 종속 변수에 미치는 영향력을 회귀식을 통해 알아보는 통계분석 방법**
- **중회귀분석 데이터의 예**

No.	악력 (kg) (x1)	신장 (cm) (x2)	체중 (kg) (x3)	공던지기 (m) (y)
1	28	146	34	22
2	46	169	57	36
3	39	160	48	24
4	25	156	38	22
5	34	161	47	27
6	29	168	50	29
7	38	154	54	26
8	23	153	40	23
9	42	160	62	31
10	27	152	39	24
11	35	155	46	23

단순회귀 분석과의 차이점 - 주의사항

- 독립변수의 문제
  - 독립변수의 유의성 : 일부 독립변수는 유의하지 않을 수도 있음
  - 독립변수의 수에 따른 기여율 (결정계수)의 증가
    - 독립변수의 수(자유도)가 많아지면 작을 때 보다 기여율이 증가함
    - 자유도를 고려한 기여율이 필요 ⓿ R^2 대신 조정 R^2 (Adjusted-R^2)을 사용

$$Adjusted - R^2 = R^2 - \frac{k(1-R^2)}{n-k-1} \quad n: \text{표본 수}, \; k = \text{독립변수의 수}$$

- 독립변수 사이의 문제
  - 하나의 독립변수가 다른 독립변수와 일정한 선형 관계가 있는 경우
    - 다중 공선성(multi-collinearity)
  - 교호작용 효과(상호작용 효과: interaction effect)
    - 하나의 독립변수 증감이 다른 변수의 증감에도 영향을 미치는 경우

- 가변수 (더미변수 : dummy variable)을 이용한 비교

독립변수 선택의 문제

- 작은 수의 독립변수로 종속변수를 설명하는 것이 많은 수의 독립변수를 사용하는 것 보다 좋음
- 종속변수와 관련없는 독립변수를 추가하더라도 R^2는 증가하는 경우가 많음
- 이론적 배경이나 사전 연구를 참조하여 선택하는 것이 최선
- 단순회귀를 이용하여 유의한 변수로 중회귀의 독립변수를 택하는 것은 잘못
- 변수 제거 전후의 R^2 값의 변화를 분석해야 함
- 독립변수 제거 시 C_p 값을 이용함

독립변수간의 영향도 비교

- 독립변수간에 종속변수에 상대적으로 많은 영향을 미치는 변수를 찾는 문제
  - 회귀계수의 크기로 비교하는 것은 옳지 않음
  - 변수간의 단위가 상이하기 때문 ➊ 비표준화회귀계수
  - 단위를 고려한 회귀계수가 필요 ➊ 표준화회귀계수
  - 표준화 회귀계수 = 비표준화회귀계수 * (S_x/S_y)
- 각 독립변수와 종속변수의 상관관계 분석
  - 부분상관관계분석을 이용
    - (Analyze -> correlate -> Partial)
    - 통제하고자 하는 변수는 'Controlling for' 에 이동

중회귀 분석과 활용의 순서

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

회귀분석의 가정

가정	설명	확인
선형성	주어진 독립변수와 종속변수 사이에 선형관계 존재	회귀식의 유의성 (F-test)
독립성	잔차들은 독립적인 관계 (자기상관없음)	Dubin-Watson test
등분산성	모든 변수의 오차는 일정	Partial regression plot
정규성	오차항은 정규분포를 따름	Normal P-P plot

회귀식의 추정과 적합성 분석

- 회귀식의 추정 방법과 적합성 분석은 단순회귀의 경우와 개념은 동일
- 주의점: 중회귀식이 의미를 가진다는 말은 중회귀식의 회귀계수 중 하나라도 의미를 지닌다 말 (회귀계수가 모두 0 만 아니면 됨)
  - $H_o : \beta_1 = \beta_2 = \ldots \beta_n = 0$ 이다.
  - $H_1 : \beta_i$ 중 하나라도 0이 아닌 것이 있다.

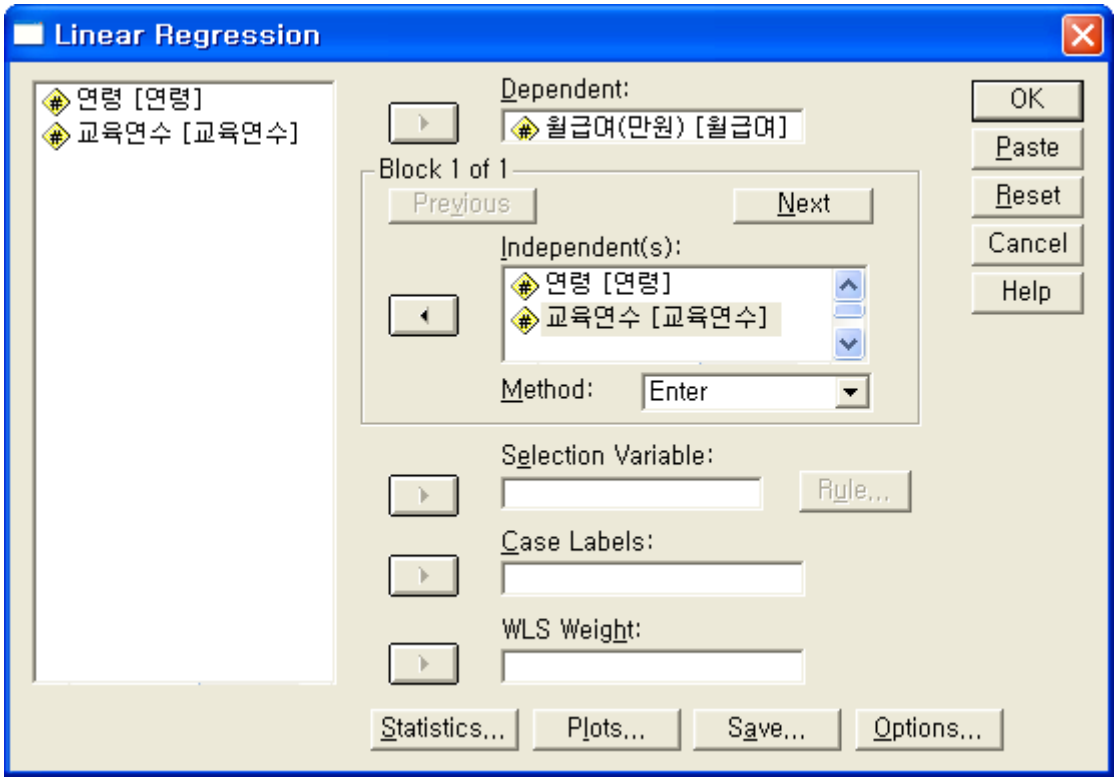
- F 통계량의 sig. 값이 ≥ 0.05이면
  - 모든 독립변수가 설명력의 증가에 기여하지 못함
  - 모든 독립변수가 통계적으로 유의미하지 않음 (모든 회귀계수가 0이므로)
  - 회귀식의 의미가 없음
- F 통계량의 sig. 값이 < 0.05이면
  - 하나 이상의 독립변수가 설명력의 증가에 기여하지 못함
  - 하나 이상의 독립변수가 통계적으로 유의미하지 않음 (0이 아닌 회귀계수가 적어도 한 개는 있으므로)
  - 회귀식의 의미가 있음 (회귀계수가 0이 아닌 독립변수로 구성하면 됨)

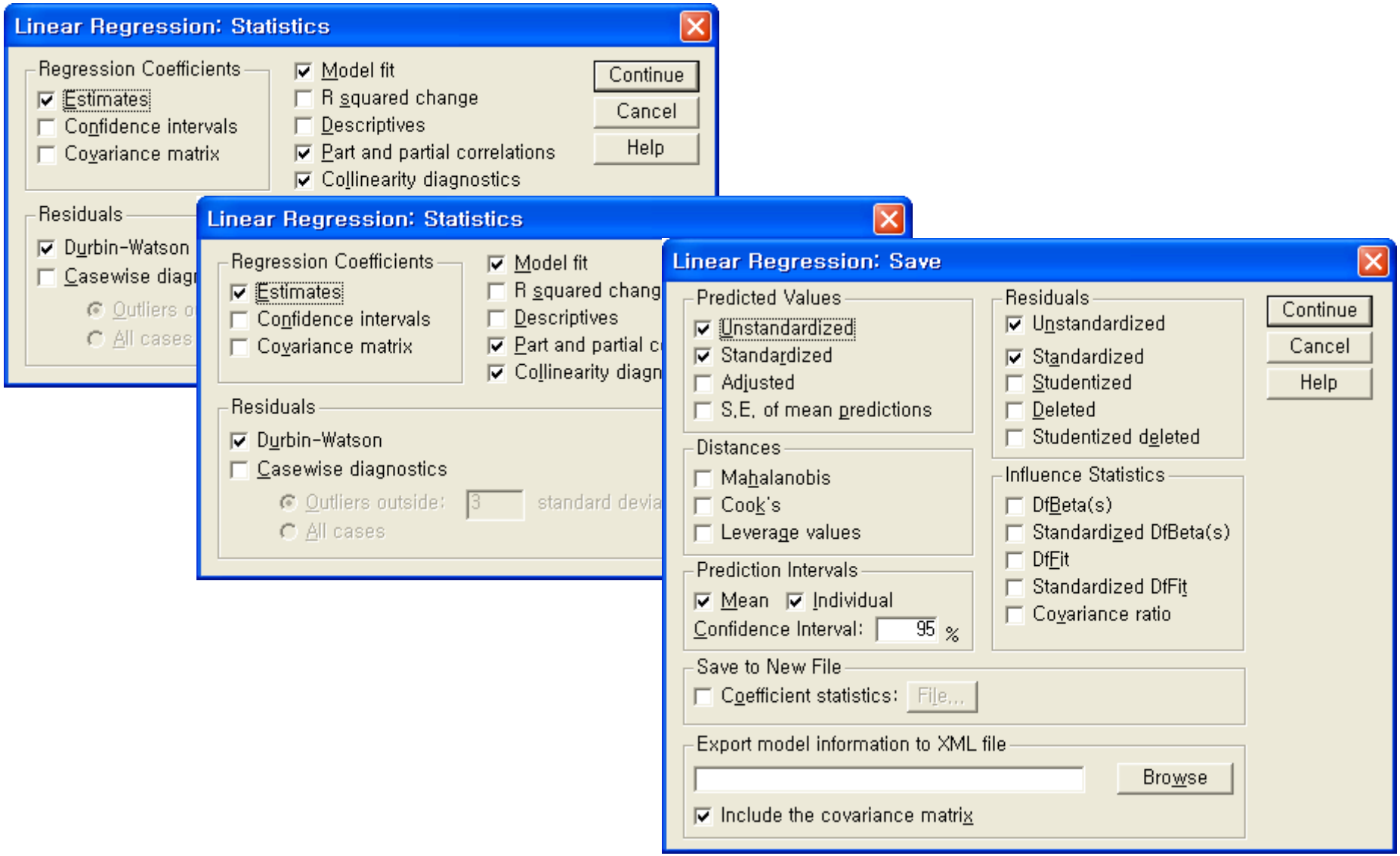
중회귀계수 유의성 분석

- 회귀식 (회귀모형)의 적합성이 있는 경우에 회귀계수의 유의성을 분석함
- 회귀식의 적합성 분석은 0이 아닌 회귀계수가 있음을 밝힘
- 어떤 회귀계수가 0이 아닌지 추가 분석이 필요 , 중회귀계수 유의성 분석을 시행
- 각 회귀계수별로 t-검정을 시행 (개별 회귀계수 값의 범위가 0을 포함하는지 검정)

실습

- **Data : 월급여.sav**




Linear 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
Linear Regression: Save
Predicted Values
Unstandardized
Standardized
Adjusted
S.E. of mean predictions
Distances
Mahalanobis
Cook's
Leverage values
Prediction Intervals
Mean
Individual
Confidence Interval:
95
%
Residuals
Unstandardized
Standardized
Studentized
Deleted
Studentized deleted
Influence Statistics
DfBeta(s)
Standardized DfBeta(s)
DfFit
Standardized DfFit
Covariance ratio
Save to New File
Coefficient statistics:
File...
Export model information to XML file
Browse
Include the covariance matrix

a Predictors: (Constant), 교육연수, 연령
b Dependent Variable: 월급여(만원)

Model Summary [b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.536[a]	.287	.253	140.971	.623

a. Predictors: (Constant), 교육연수, 연령

b. Dependent Variable: 월급여(만원)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	35678.692	2	167839.346	8.446	.001[a]
	Residual	34662.669	42	19872.921		
	Total	1170341.4	44			

a. Predictors: (Constant), 교육연수, 연령

b. Dependent Variable: 월급여(만원)

Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-241.437	123.070		-1.962	.056
	연령	4.611	1.551	.405	2.973	.005
	교육연수	23.454	6.562	.486	3.574	.001

a. Dependent Variable: 월급여(만원)

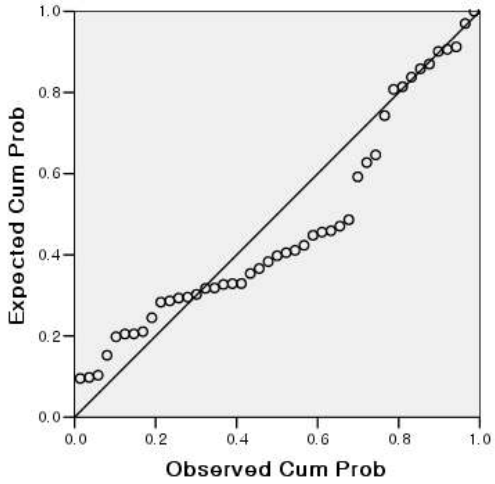
Correlations			Collinearity Statistics	
Zero-order	Partial	Part	Tolerance	VIF
.264	.417	.387	.917	1.091
.370	.483	.466	.917	1.091

Collinearity Diagnostics [a]

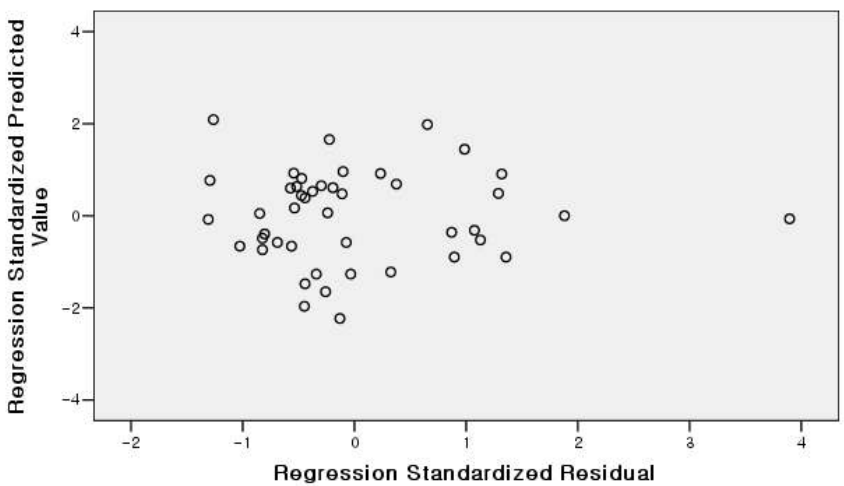
Model	Dimension	Eigenvalue	Condition Index	Variance Proportions		
				(Constant)	연령	교육연수
1	1	2.876	1.000	.00	.01	.01
	2	.104	5.247	.00	.48	.24
	3	.019	12.294	.99	.51	.75

a. Dependent Variable: 월급여(만원)

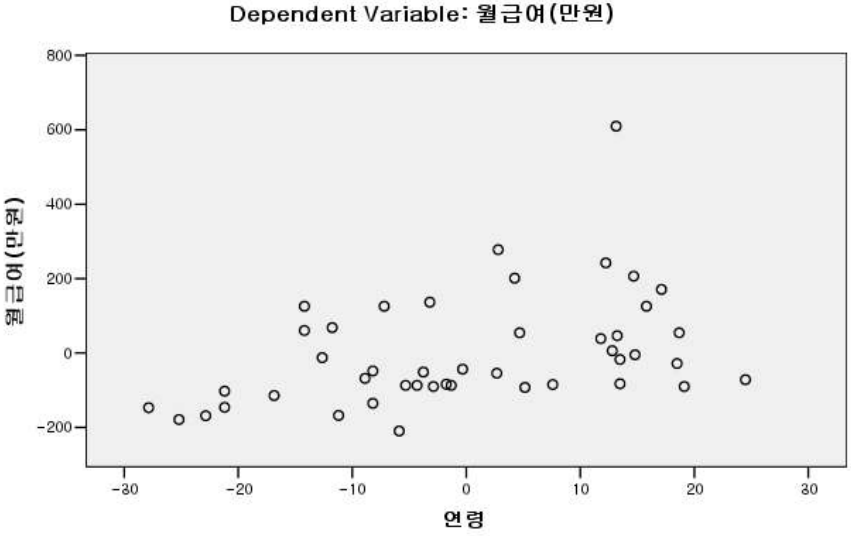
Normal P-P Plot of Regression Standardized Residual



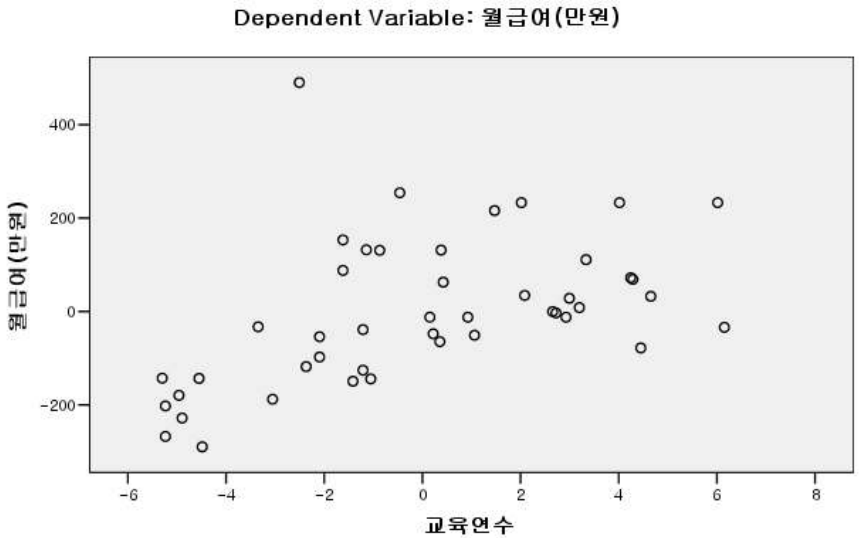
Scatterplot



Partial Regression Plot



Partial Regression Plot



가변수(Dummy variable)를 통한 분석

- **가변수(Dummy variable) : 범주형의 변수를 회귀분석에 사용하기 위해 사용되는 변수**
- **독립변수 가운데 질적인 변수가 있는 경우**
  - **예) 성별에 따른 근무년수와 임금의 관계 분석**
- **가변수는 분석의 편의를 위해 0 또는 1의 값을 부여**
  - **다른 값의 부여도 가능하나 결과해석이 어려움**
- **가변수의 개수는 (질적인 변수가 가지는 값의 개수 -1)**
- **성별이란 변수는 남/녀 (0/1)의 두 값을 가짐. 따라서 가변수의 개수는 (2-1)개**
- **학년 이라는 변수는 1/2/3/4의 4가지 값을 가짐, 가변수는 (4-1)개**

실습

- **Data : 성별_임금.sav**
- **Dummy variable = 성별 (남: 0, 여: 1)**

Model Summary [b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.415[a]	.172	.172	48.039

a. Predictors: (Constant), 성별, 교육연수(년)

b. Dependent Variable: 임금(만원)

ANOVA [b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1526081.3	2	763040.634	330.648	.000[a]
	Residual	7338520.4	3180	2307.711		
	Total	8864601.7	3182			

a. Predictors: (Constant), 성별, 교육연수(년)

b. Dependent Variable: 임금(만원)

Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	70.626	4.215		16.757	.000	62.362	78.890
	교육연수(년)	5.230	.305	.277	17.134	.000	4.632	5.829
	성별	-35.917	1.751	-.332	-20.516	.000	-39.350	-32.485

a. Dependent Variable: 임금(만원)

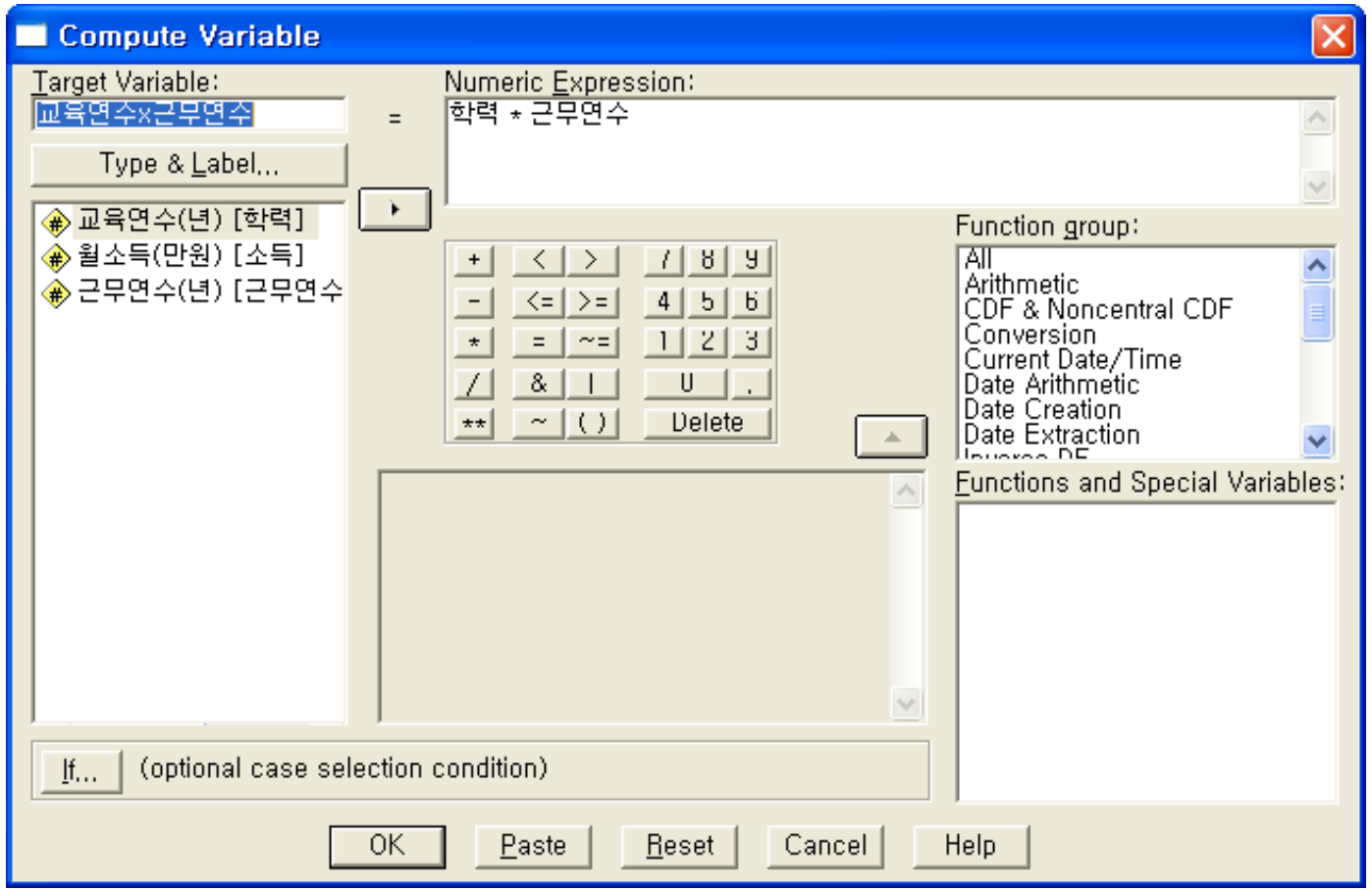
- 회귀식은?
- 남여간 임금차이는?

상호작용(교호작용 : interaction)

- 상호작용이란 : 독립변수들 간의 영향을 주고 받아 종속변수에 미치는 영향력이 서로 달라지는 경우를 말함
- 상호작용 항을 추가 (대부분 상호작용을 하는 독립변수의 곱으로 이루어진 항)
- 예) 교육연수가 더 많은 사람이 근무연수에 따른 임금이 더 많은 경우
  - 임금 = $a+b_1$*교육연수 + b_2*근무연수 + b_3*교육연수*근무연수
- 상호작용이 있는 경우의 회귀식의 해석
  - 상호작용항이 통계적으로 의미가 있는지 확인 (상호작용항이 유의미하면 개별 독립변수들의 유의도는 무시)
  - 상호작용항이 무의미하면 상호작용항을 제거하고 개별 독립변수만으로 이루어진 회귀식을 사용

실습

- **Data : 상호작용.sav**
- **실습과정 : 교육연수(학력)과 근무연수의 곱으로 된 상호작용항을 추가 후 회귀분석**



상호작용

Model Summary [b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.361[a]	.131	.128	45.829	1.578

a. Predictors: (Constant), 교육연수x근무연수, 교육연수(년), 근무연수(년)

b. Dependent Variable: 월소득(만원)

ANOVA [b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	91455.366	3	130485.122	62.127	.000[a]
	Residual	2606476.0	1241	2100.303		
	Total	2997931.3	1244			

a. Predictors: (Constant), 교육연수x근무연수, 교육연수(년), 근무연수(년)

b. Dependent Variable: 월소득(만원)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	145.577	19.874		7.325	.000
	교육연수(년)	2.646	1.464	.149	1.807	.071
	근무연수(년)	-5.791	3.033	-.244	-1.909	.056
	교육연수x근무연수	.619	.229	.372	2.702	.007

a. Dependent Variable: 월소득(만원)

95% Confidence Interval for B		Correlations			Collinearity Statistics	
Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
106.587	184.568					
-.227	5.519	.342	.051	.048	.104	9.655
-11.742	.160	.030	-.054	-.051	.043	23.323
.170	1.069	.234	.076	.072	.037	27.005

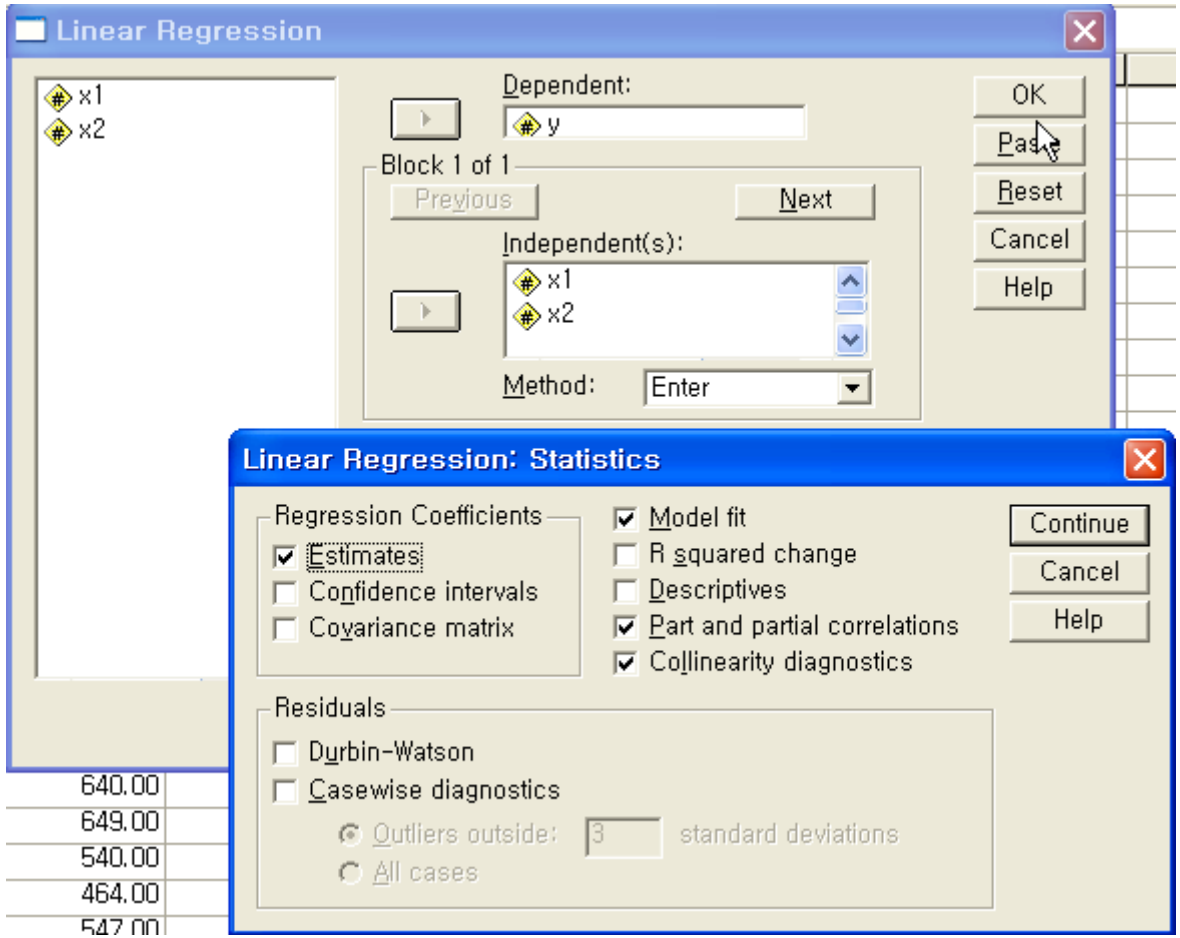
다중공선성

- 다중공선성이란 : 독립변수들 간에 서로 밀접한 관계가 있는 경우
  - ➊ 정상적인 회귀분석 결과가 나오지 않음
- 독립변수들간 완벽한 선형관계 : 극단적 다중공선성
- 독립변수들간 선형에 가까운 관계 : 준 극단적 다중공선성

- 다중공선성의 진단
  - 허용값(tolerance)가 0.4 이하 이거나 분산팽창인자 (VIF)가 10 이상
    - 변수들간의 다중공선성이 존재한다는 근거, 구체적인 변수는 모름
  - 개별 변수의 상태지수 (Condition index)가 15 이상이거나 하나의 고유근(eigen value)가 가지는 분산비율들 (variance proportion)이 공통적으로 높은 경우
- 해결책
  - 공선성이 높은 변수들이 유사한 개념을 측정하고 있다면 하나를 제거
  - 서로 다른 개념을 측정하고 있다면…. ?

실습

- **Data : 다중공선성.sav**



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.704[a]	.496	.474	81.16582

a. Predictors: (Constant), x2, x1

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	91911.393	2	145955.696	22.155	.000[a]
	Residual	96455.086	45	6587.891		
	Total	88366.479	47			

a. Predictors: (Constant), x2, x1

b. Dependent Variable: y

Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-249.706	125.282		-1.993	.052
	x1	-20.594	42.202	-1.021	-.488	.628
	x2	35.093	42.635	1.722	.823	.415

a. Dependent Variable: y

Correlations			Collinearity Statistics	
Zero-order	Partial	Part	Tolerance	VIF
.699	-.073	-.052	.003	390.972
.702	.122	.087	.003	390.972

Collinearity Diagnostics [a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions		
				(Constant)	x1	x2
1	1	2.994	1.000	.00	.00	.00
	2	.006	22.214	.96	.00	.00
	3	1.16E-005	507.434	.04	1.00	1.00

a. Dependent Variable: y