회귀분석론 과제 4 오유진 (2015325)

8.1

8.1.1

CORR 프로시저

4 개의 변수:	x1 x2 x3 x4

단순 통계량

변수	N	평균	표준편차	합	최솟값	최댓값
x1	13	7.46154	5.88239	97.00000	1.00000	21.00000
x2	13	48.15385	15.56088	626.00000	26.00000	71.00000
x3	13	11.76923	6.40513	153.00000	4.00000	23.00000
x4	13	30.00000	16.73818	390.00000	6.00000	60.00000

피어슨 상관 계수, N = 13
H0: Rho=0 가정하에서 Prob > |r|

	x1	x2	x3	x4
x1	1.00000	0.22858 0.4526	-0.82413 0.0005	-0.24545 0.4189
x2	0.22858 0.4526	1.00000	-0.13924 0.6501	-0.97295 <.0001
x3	-0.82413 0.0005	-0.13924 0.6501	1.00000	0.02954 0.9237
x4	-0.24545 0.4189	-0.97295 <.0001	0.02954 0.9237	1.00000

x2와 x4, x1와 x3의 표본 상관계수가 -1에 근접하기 때문에 강한 음의 선형관계를 보이고 있다.

8.1.2

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	62.40537	70.07096	0.89	0.3991	0
x1	1	1.55110	0.74477	2.08	0.0708	38.49621
x2	1	0.51017	0.72379	0.70	0.5009	254.42317
x3	1	0.10191	0.75471	0.14	0.8959	46.86839
x4	1	-0.14406	0.70905	-0.20	0.8441	282.51286

VIF 최댓값이 282,254로 다중공산성 문제가 있다고 볼 수 있다.

8.1.3

Collinearity Diagnostics (intercept adjusted)

Number	Eigenvalue	Condition Index	Proportion of Variation			
			x1	x2	x3	x4
1	2.23570	1.00000	0.00263	0.00055897	0.00148	0.00047533
2	1.57607	1.19102	0.00427	0.00042729	0.00495	0.00045729
3	0.18661	3.46134	0.06352	0.00208	0.04650	0.00072440
4	0.00162	37.10634	0.92958	0.99693	0.94707	0.99834

고유값은 eigenvalue, 상태지표는 condition index이다. 상태지표 중 가장 큰 수는 37.10634로 상태수이다. 또한 상태수는 30보다 크므로 다중공산성이 존재한다고 할 수 있다. 또한 상태지표 중 30보다 큰 수는 한 개이므로 설명변수 사이에 한 개의 선형관계가 존재한다고 볼 수 있다.

8.1.4

Collinearity Diagnostics (intercept adjusted)						
Number	Eigenvalue	Condition Index	Proportion of Variation			
			x1	x2	x3	x4
1	2.23570	1.00000	0.00263	0.00055897	0.00148	0.00047533
2	1.57607	1.19102	0.00427	0.00042729	0.00495	0.00045729
3	0.18661	3.46134	0.06352	0.00208	0.04650	0.00072440
4	0.00162	37.10634	0.92958	0.99693	0.94707	0.99834

분산 비율은 proportion of variation으로 위의 표와 같다.
분산비율을 보면 π41, π42, π43, π44가 다 0.9이상으로 x1, x2, x3, x4의 다중 공산성이 존재한다고 볼 수 있다.

결론: 따라서 8.1.2, 8.1.3, 8.1.4에 의해서 X1, X2, X3, X4의 다중공산성이 존재한다고 볼 수 있다.

8.4

8.4.1

Number in Model	R-Square	Adjusted R-Square	C(p)	MSE	Variables in Model
1	0.6745	0.6450	138.7308	80.35154	x4
1	0.6663	0.6359	142.4864	82.39421	x2
1	0.5339	0.4916	202.5488	115.06243	x1
1	0.2859	0.2210	315.1543	176.30913	x3
2	0.9787	0.9744	2.6782	5.79045	x1 x2
2	0.9725	0.9670	5.4959	7.47621	x1 x4
2	0.9353	0.9223	22.3731	17.57380	x3 x4
2	0.8470	0.8164	62.4377	41.54427	x2 x3
2	0.6801	0.6161	138.2259	86.88801	x2 x4
2	0.5482	0.4578	198.0947	122.70721	x1 x3
3	0.9823	0.9764	3.0182	5.33030	x1 x2 x4
3	0.9823	0.9764	3.0413	5.34562	x1 x2 x3
3	0.9813	0.9750	3.4968	5.64846	x1 x3 x4
3	0.9728	0.9638	7.3375	8.20162	x2 x3 x4
4	0.9824	0.9736	5.0000	5.98295	x1 x2 x3 x4

adjusted R-square를 기준으로 보면 adjusted R-square가 가장 작은 값인 p=1일 때, x3이 포함된 모형이 최적모형이다.

8.4.2.

Number in Model	R-Square	Adjusted R-Square	C(p)	MSE	Variables in Model
1	0.6745	0.6450	138.7308	80.35154	x4
1	0.6663	0.6359	142.4864	82.39421	x2
1	0.5339	0.4916	202.5488	115.06243	x1
1	0.2859	0.2210	315.1543	176.30913	x3
2	0.9787	0.9744	2.6782	5.79045	x1 x2
2	0.9725	0.9670	5.4959	7.47621	x1 x4
2	0.9353	0.9223	22.3731	17.57380	x3 x4
2	0.8470	0.8164	62.4377	41.54427	x2 x3
2	0.6801	0.6161	138.2259	86.88801	x2 x4
2	0.5482	0.4578	198.0947	122.70721	x1 x3
3	0.9823	0.9764	3.0182	5.33030	x1 x2 x4
3	0.9823	0.9764	3.0413	5.34562	x1 x2 x3
3	0.9813	0.9750	3.4968	5.64846	x1 x3 x4
3	0.9728	0.9638	7.3375	8.20162	x2 x3 x4
4	0.9824	0.9736	5.0000	5.98295	x1 x2 x3 x4

Cp를 기준으로 보면, 가장 작은 값인 P=2일 때, x1, x2가 포함된 모형이 최적모형이다.

8.4.4.

Forward Selection: Step 1

Variable x4 Entered: R-Square = 0.6745 and C(p) = 138.7308

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1831.89616	1831.89616	22.80	0.0006
Error	11	883.86692	80.35154		
Corrected Total	12	2715.76308			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	117.56793	5.26221	40108	499.16	<.0001
x4	-0.73816	0.15460	1831.89616	22.80	0.0006

부분 F-검정통계량이 가장 큰 설명변수는 x4이고 p값이 매우 작으므로 유의하므로 x4를 설명변수에 포함한다.

Forward Selection: Step 2

Variable x1 Entered: R-Square = 0.9725 and C(p) = 5.4959

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2641.00096	1320.50048	176.63	<.0001
Error	10	74.76211	7.47621		
Corrected Total	12	2715.76308			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	103.09738	2.12398	17615	2356.10	<.0001
x1	1.43996	0.13842	809.10480	108.22	<.0001
x4	-0.61395	0.04864	1190.92464	159.30	<.0001

Bounds on condition number: 1.0641, 4.2564

x4에 포함된 모형에 step1과 동일과정으로 나머지 설명변수를 추가한다. 여기선 x1을 추가한다.

Forward Selection: Step 3

Variable x2 Entered: R-Square = 0.9823 and C(p) = 3.0182

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2667.79035	889.26345	166.83	<.0001
Error	9	47.97273	5.33030		
Corrected Total	12	2715.76308			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	71.64831	14.14239	136.81003	25.67	0.0007
x1	1.45194	0.11700	820.90740	154.01	<.0001
x2	0.41611	0.18561	26.78938	5.03	0.0517
x4	-0.23654	0.17329	9.93175	1.86	0.2054

Bounds on condition number: 18.94, 116.36

No other variable met the 0.2000 significance level for entry into the model.

위와 같은 방법으로 x2의 p값은 유의수준 0.2보다 작으므로 설명변수 x2를 추가한다.

따라서 전진선택법을 이용하여 구한 최적모형은 설명변수 x1, x2, x4를 포함한 모형으로 다음과 같다.

Summary of Forward Selection

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x4	1	0.6745	0.6745	138.731	22.80	0.0006
2	x1	2	0.2979	0.9725	5.4959	108.22	<.0001
3	x2	3	0.0099	0.9823	3.0182	5.03	0.0517

8.4.5

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.9824 and C(p) = 5.0000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	2667.89944	666.97486	111.48	<.0001
Error	8	47.86364	5.98295		
Corrected Total	12	2715.76308			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	62.40537	70.07096	4.74552	0.79	0.3991
x1	1.55110	0.74477	25.95091	4.34	0.0708
x2	0.51017	0.72379	2.97248	0.50	0.5009
x3	0.10191	0.75471	0.10909	0.02	0.8959
x4	-0.14406	0.70905	0.24697	0.04	0.8441

Bounds on condition number: 282.51, 2489.2

4개의 설명변수 모드 포함하는 다중회귀모형을 적합하면 왼쪽과 같다. 이 중 부분 F-검정통계량이 가장 작은 x3을 선택해 유의성검정을 실시한다. x3의 p값은 매우 크므로 x3을 제거한다.

Backward Elimination: Step 1

Variable x3 Removed: R-Square = 0.9823 and C(p) = 3.0182

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2667.79035	889.26345	166.83	<.0001
Error	9	47.97273	5.33030		
Corrected Total	12	2715.76308			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	71.64831	14.14239	136.81003	25.67	0.0007
x1	1.45194	0.11700	820.90740	154.01	<.0001
x2	0.41611	0.18561	26.78938	5.03	0.0517
x4	-0.23654	0.17329	9.93175	1.86	0.2054

Bounds on condition number: 18.94, 116.36

x3을 제거하고 나머지 설명변수를 포함한 모형을 적합하면 왼쪽과 같다. 부분 F-검정통계량이 가장 작은 x4는 p값이 유의수준 0.1다 크므로 x4를 제거할 수 있다.

Backward Elimination: Step 2

Variable x4 Removed: R-Square = 0.9787 and C(p) = 2.6782

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2657.85859	1328.92930	229.50	<.0001
Error	10	57.90448	5.79045		
Corrected Total	12	2715.76308			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	52.57735	2.28617	3062.60416	528.91	<.0001
x1	1.46831	0.12130	848.43186	146.52	<.0001
x2	0.66225	0.04585	1207.78227	208.58	<.0001

Bounds on condition number: 1.0551, 4.2205

All variables left in the model are significant at the 0.1000 level.

x4를 제거하고 나머지 설명변수 x1, x2를 회귀시키면 왼쪽과 같다. 두 설명변수모두 부분 F-검정통계량도 크고 p값도 매우 작으므로 위 과정을 중단해도 된다고 판단할 수 있다. 따라서 x1, x2를 포함한 모형이 최적모형이 된다.

따라서 후진제거법을 이용하여 구한 최적모형은 설명변수 x3, x4를 제거한 설명변수 x1, x2를 포함한 모형으로 다음과 같다.

Summary of Backward Elimination

Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x3	3	0.0000	0.9823	3.0182	0.02	0.8959
2	x4	2	0.0037	0.9787	2.6782	1.86	0.2054

8.4.6.

Stepwise Selection: Step 1

Variable x4 Entered: R-Square = 0.6745 and C(p) = 138.7308

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1831.89616	1831.89616	22.80	0.0006
Error	11	883.86692	80.35154		
Corrected Total	12	2715.76308			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	117.56793	5.26221	40108	499.16	<.0001
x4	-0.73816	0.15460	1831.89616	22.80	0.0006

Bounds on condition number: 1, 1

단순회귀에서 F-검정통계량이 가장 큰 설명변수 x4는 유의수준이 유의수준 0.20보다 작으므로 회귀모형에 포함시킨다.

Stepwise Selection: Step 2

Variable x1 Entered: R-Square = 0.9725 and C(p) = 5.4959

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2641.00096	1320.50048	176.63	<.0001
Error	10	74.76211	7.47621		
Corrected Total	12	2715.76308			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	103.09738	2.12398	17615	2356.10	<.0001
x1	1.43996	0.13842	809.10480	108.22	<.0001
x4	-0.61395	0.04864	1190.92464	159.30	<.0001

Bounds on condition number: 1.0641, 4.2564

위의 모형에 부분 F-검정통계량이 두번째 큰 설명변수 x1을 선택한다. 유의성검정을 해보면 x1의 p값은 매우 작으므로 설명변수 x1을 회귀모형에 포함시킨다.

Stepwise Selection: Step 3

Variable x2 Entered: R-Square = 0.9823 and C(p) = 3.0182

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2667.79035	889.26345	166.83	<.0001
Error	9	47.97273	5.33030		
Corrected Total	12	2715.76308			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	71.64831	14.14239	136.81003	25.67	0.0007
x1	1.45194	0.11700	820.90740	154.01	<.0001
x2	0.41611	0.18561	26.78938	5.03	0.0517
x4	-0.23654	0.17329	9.93175	1.86	0.2054

Bounds on condition number: 18.94, 116.36

위와 같은 방법으로 설명변수 x2를 선택한다. x2의 유의수준은 0.1보다 작으므로 설명변수 x2를 회귀모형에 포함시킨다.

Stepwise Selection: Step 4

Variable x4 Removed: R-Square = 0.9787 and C(p) = 2.6782

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2657.85859	1328.92930	229.50	<.0001
Error	10	57.90448	5.79045		
Corrected Total	12	2715.76308			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	52.57735	2.28617	3062.60416	528.91	<.0001
x1	1.46831	0.12130	848.43186	146.52	<.0001
x2	0.66225	0.04585	1207.78227	208.58	<.0001

Bounds on condition number: 1.0551, 4.2205

All variables left in the model are significant at the 0.1000 level.

No other variable met the 0.2000 significance level for entry into the model.

위 모형에서 부분 F-검정 통계량이 가장 작은 설명변수 x4를 유의성검정을 실시한다. x4의 p값은 유의수준 0.10보다 크므로 설명변수 x4를 제거할 수 있다.

x1, x2는 F-검정통계량 값도 m고 p값도 매우 작으므로 이 과정을 중단해도 된다고 판단된다.

따라서 단계적 회귀방법을 통해 구한 최적모형은 다음과 같다.

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x4		1	0.6745	0.6745	138.731	22.80	0.0006
2	x1		2	0.2979	0.9725	5.4959	108.22	<.0001
3	x2		3	0.0099	0.9823	3.0182	5.03	0.0517
4		x4	2	0.0037	0.9787	2.6782	1.86	0.2054

8.4.7.

위에 따르면 전진제거법, 단계별 회귀방법을 통해 구한 최적모형은 설명변수 x1, x2를 포함한 회귀모형이고 후진제거법을 통해 구한 최적모형은 설명변수 x1, x2 x4를 포함한 회귀모형이다. 후진제거법을 통해 구한 최적모형이 전진 제거법과 단계별 회귀방법으로 구한 최적모형보다 더 편리하고 간단하다.