회귀분석
팀프로젝트

교통사고 원인 분석 보고서

기한정 1614076

박유희 1611888

이도진 1514606

목차

1. 분석설명 & 데이터 탐색

<분석 목적>

교통사고 횟수에 영향을 미치는 교통사고의 원인을 알아보고, 도로의 종류에 따라 서로 다른 원인의 영향을 받는지 혹은 상관이 없는지를 알아보기 위하여 분석을 시작하게 되었음.

<자료설명(n=39)>

1973년 미국 미네소타주의 도로 39개 구간에서 자동차 주행거리에 대한 자료로, 100만 마일당 사고의 횟수를 반응변수로 하고 사고의 원인이 될 수 있을 것으로 예상되는 설명변수 13개에 대한 조사결과이다.

<변수설명>

Y: 100만 마일의 자동차 주행거리당 사고의 횟수
X1 : 구간의 길이(miles)
X2: 하루 평균 통과 자동차의 수(1000대)
X3: 트럭의 비율
X4: 제한속도(mile/hour)
X5: 차선의 폭(feet)
X6: 갓길의 폭(feet)
X7: 1마일당 고속도로 진입로의 수
X8: 1마일당 신호등이 있는 교차로의 수
X9: 1마일당 진입로의 수
X10: 차선의 수
X11: 1:연방고속도로, 0:기타
X12: 1:주고속도로, 0:기타
X13: 1:주간선도로,0:기타

X1-X10: 교통사고 원인에 관한 변수
X11-X13: 도로의 종류를 나타내는 지시변수

<단순통계량>

단순 통계량

변수	N	평균	표준편차	합	최솟값	최댓값
y	39	3.93333	1.98604	153.40000	1.61000	9.23000
x1	39	12.88410	7.60968	502.48000	2.96000	40.09000
x2	39	19.61538	18.61185	765.00000	1.00000	73.00000
x3	39	9.33333	2.35454	364.00000	6.00000	15.00000
x4	39	55.00000	5.84898	2145	40.00000	70.00000
x5	39	11.94872	0.45588	466.00000	10.00000	13.00000
x6	39	6.87179	3.03644	268.00000	1.00000	10.00000
x7	39	0.29641	0.41117	11.56000	0	1.54000
x8	39	0.40051	0.63339	15.62000	0	2.51000
x9	39	12.15897	9.31834	474.20000	2.20000	53.00000
x10	39	3.12821	1.36072	122.00000	2.00000	8.00000

x2의 경우 최솟값에 대한 최댓값의 비가 73으로 매우 높았다. 하루 통과 자동차의 수의 차이가 도로별로 매우 크다는 것을 알게 되었다. 또한 최댓값/최솟값이 10이상인 변수는 x1, x6, x9였고, x7과 x8은 0을 포함한다. 결측은 존재하지 않았다.

<상관계수>

피어슨 상관 계수, N = 39
H0: Rho=0 가정하에서 Prob > |r|

	y	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
y	1.00000	-0.46529 0.0028	-0.02857 0.8629	-0.51252 0.0009	-0.68098 <.0001	-0.00562 0.9729	-0.38691 0.0150	-0.02484 0.8807	0.56448 0.0002	0.75203 <.0001	-0.03298 0.8420
x1	-0.46529 0.0028	1.00000	-0.27157 0.0945	0.49594 0.0013	0.18624 0.2563	-0.31065 0.0543	-0.10493 0.5250	-0.24756 0.1286	-0.32179 0.0458	-0.23871 0.1433	-0.20250 0.2163
x2	-0.02857 0.8629	-0.27157 0.0945	1.00000	-0.09688 0.5582	0.24416 0.1342	0.12788 0.4379	0.45731 0.0034	0.90370 <.0001	0.14547 0.3769	-0.22398 0.1705	0.82393 <.0001
x3	-0.51252 0.0009	0.49594 0.0013	-0.09688 0.5582	1.00000	0.29618 0.0671	-0.15527 0.3452	0.00613 0.9704	-0.06723 0.6843	-0.45026 0.0040	-0.36027 0.0243	-0.15332 0.3514
x4	-0.68098 <.0001	0.18624 0.2563	0.24416 0.1342	0.29618 0.0671	1.00000	0.09869 0.5500	0.68901 <.0001	0.24128 0.1389	-0.41022 0.0095	-0.68152 <.0001	0.26452 0.1037
x5	-0.00562 0.9729	-0.31065 0.0543	0.12788 0.4379	-0.15527 0.3452	0.09869 0.5500	1.00000	-0.04290 0.7954	0.10288 0.5331	0.04202 0.7995	-0.04201 0.7995	0.09572 0.5621
x6	-0.38691 0.0150	-0.10493 0.5250	0.45731 0.0034	0.00613 0.9704	0.68901 <.0001	-0.04290 0.7954	1.00000	0.37502 0.0187	-0.13406 0.4159	-0.42495 0.0070	0.48177 0.0019
x7	-0.02484 0.8807	-0.24756 0.1286	0.90370 <.0001	-0.06723 0.6843	0.24128 0.1389	0.10288 0.5331	0.37502 0.0187	1.00000	0.06951 0.6741	-0.20016 0.2218	0.69791 <.0001
x8	0.56448 0.0002	-0.32179 0.0458	0.14547 0.3769	-0.45026 0.0040	-0.41022 0.0095	0.04202 0.7995	-0.13406 0.4159	0.06951 0.6741	1.00000	0.49869 0.0012	0.24969 0.1253
x9	0.75203 <.0001	-0.23871 0.1433	-0.22398 0.1705	-0.36027 0.0243	-0.68152 <.0001	-0.04201 0.7995	-0.42495 0.0070	-0.20016 0.2218	0.49869 0.0012	1.00000	-0.20878 0.2021
x10	-0.03298 0.8420	-0.20250 0.2163	0.82393 <.0001	-0.15332 0.3514	0.26452 0.1037	0.09572 0.5621	0.48177 0.0019	0.69791 <.0001	0.24969 0.1253	-0.20878 0.2021	1.00000

상관계수 값이 높은값은 (y,x4) (y,x9) (x2,x7) (x2,x10) (x4,x6) (x4, x9) (x7,x10)으로 유의확률값이 0.0001값보다 작다고 나왔다. 반응변수는 제한속도(x4)와 1마일당 진입로의 수(x9)와 강한 상관관계가 있었다.

설명변수끼리의 상관관계를 보면, 하루 평균 통과 자동차의 수(x2)는 고속도로 진입로의 수(x7)과 차선의수(x10)과 강한 상관관계를 보였고, 제한속도(x4)와 갓길의 폭(x6), 1마일당 진입로의 수(x9), 1마일당 고속도로 진입로의 수(x7)와 차선의 수(x10) 간에 강한 상관관계가 있었다.

<자료진단>

본 분석 전에 모든 설명변수를 포함하여 단순회귀를 적합하였을 때 자료진단결과이다.

OBS	y	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	i1	i2	i3	e1	r1	cd	t1
23	2.83	11.39	5	9	50	12	8	0.00	0.09	11.1	2	0	1	0	-1.33431	-1.30580	0.04568	-1.32542
24	1.81	22.00	5	15	60	12	7	0.00	0.00	6.8	2	0	1	0	0.53605	0.52601	0.00756	0.51826
25	9.23	3.58	23	6	40	12	2	0.56	2.51	53.0	4	0	0	1	-1.58678	-2.40246	0.94461	-2.68406
26	8.60	3.23	13	6	45	12	2	0.31	0.93	17.3	2	0	0	1	1.77830	1.76033	0.09006	1.84275
27	8.21	7.73	7	8	55	12	8	0.13	0.52	27.3	2	0	0	1	2.54998	2.44085	0.13426	2.74024
28	2.93	14.41	10	10	55	12	6	0.00	0.07	18.0	2	0	0	1	-1.17824	-1.06704	0.01444	-1.07013

- max|ti| = 2.74024 < t(0.05; n=39, p'=14)=3.65
- max|Di|= 0.94461으로 1에 가깝다.

따라서 이상점값은 없고, 25번째 관측치는 영향력이 큰 값이라는 판단을 내렸다.

2. 모형진단 후 변수변환

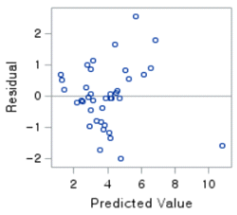
<총괄분석>

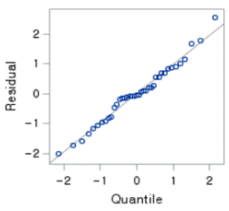
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	113.99240	8.76865	6.11	<.0001
Error	25	35.89367	1.43575		
Corrected Total	38	149.88607			

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	13.65821	6.87272	1.99	0.0579
x1	1	-0.06475	0.03337	-1.94	0.0637
x2	1	-0.00404	0.03394	-0.12	0.9063
x3	1	-0.10015	0.11473	-0.87	0.3910
x4	1	-0.12378	0.08168	-1.52	0.1422
x5	1	-0.13381	0.59792	-0.22	0.8247
x6	1	0.01411	0.16217	0.09	0.9313
x7	1	-0.47548	1.28274	-0.37	0.7140
x8	1	0.71364	0.52521	1.36	0.1864
x9	1	0.06659	0.04257	1.56	0.1303
x10	1	0.02668	0.28383	0.09	0.9259
i1	1	0.54359	1.72827	0.31	0.7557
i2	1	-1.00978	1.10561	-0.91	0.3698
i3	1	-0.54802	0.97562	-0.56	0.5793

분산분석표 참고하면 F값이 6.11로 귀무가설을 기각하므로 베타는 모두 0이라는 귀무가설 기각합니다, 따라서 적어도 하나의 베타는 0이 아니며 적어도 하나의 설명변수는 반응변수와 선형의 상관관계가 있음을 알 수 있었습니다. 총괄분석 결과, 다른 모든 변수가 있을 때 각각 하나씩 제거해도 무방한 변수들이 여럿 보였습니다.

<모형진단>

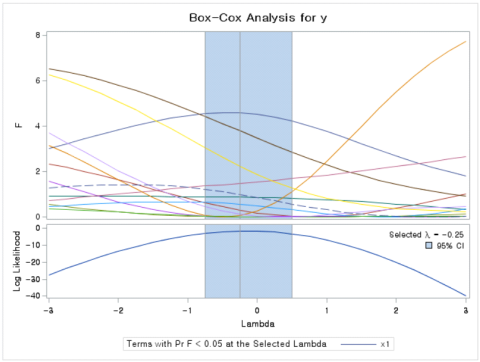




정규성 검정				
검정	통계량		p 값	
Shapiro-Wilk	W	0.982347	Pr < W	0.7873
Kolmogorov-Smirnov	D	0.117741	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.061543	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.341711	Pr > A-Sq	>0.2500

모형진단 결과, 잔차산점도를 살펴보면 x축이 증가함에 따라 y축의 값의 퍼짐의 정도가 증가하고 있습니다. 따라서 선형성과 등분산성에 문제가 있음을 알 수 있었습니다. 정규확률그림을 보면 직선에서 많이 벗어나지 않으며, w-통계량의 값이 0.98로 유의수준 0.05에서의 정규성을 따른다는 귀무가설을 기각하지 못하므로 정규성을 만족합니다.

<반응변수 변수변환(Box-Cox)>



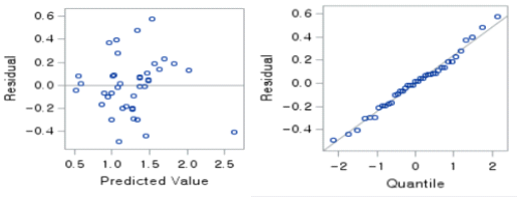
선형성과 등분산성을 만족시키기 위해 반응변수 변환을 시도했습니다. Box-Cox 변환 결과 λ를 -0.25로 추정하는데 0에 가까우므로 편의상 y를 log(y)로 변환했습니다.

<반응변수 변환 후 총괄분석>

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.08391	1.70319	2.40	0.0243
x1	1	-0.01756	0.00827	-2.12	0.0437
x2	1	0.00357	0.00841	0.42	0.6753
x3	1	-0.02640	0.02843	-0.93	0.3619
x4	1	-0.03772	0.02024	-1.86	0.0742
x5	1	-0.01249	0.14818	-0.08	0.9335
x6	1	0.00270	0.04019	0.07	0.9470
x7	1	-0.22258	0.31789	-0.70	0.4903
x8	1	0.16222	0.13016	1.25	0.2242
x9	1	0.00550	0.01055	0.52	0.6068
x10	1	-0.00404	0.07034	-0.06	0.9546
i1	1	-0.08427	0.42830	-0.20	0.8456
i2	1	-0.37572	0.27399	-1.37	0.1825
i3	1	-0.22431	0.24178	-0.93	0.3624

y를 log(y)로 반응변수 변환 후 총괄분석 결과, 다른 모든 변수가 있을 때 각각 하나씩 제거해도 무방한 변수들이 여럿 보였습니다.

<반응변수 변환 후 모형진단>



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3.80098	3.80098	1.95	0.1707
Error	37	72.06262	1.94764		
Corrected Total	38	75.86360			

정규성 검정				
검정	통계량		p 값	
Shapiro-Wilk	W	0.979756	Pr < W	0.6947
Kolmogorov-Smirnov	D	0.081011	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.040293	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.280816	Pr > A-Sq	>0.2500

반응변수 변환 후 모형진단 결과, 변환 후 잔차산점도와 변환 전 잔차산점도를 비교하면 선형성과 등분산성의 문제가 많이 해결되었음을 알 수 있었습니다.

스코어검정결과, u=(e**2)/(2.20439/39),

SSreg=3.80098,

S= 3.80098/2 = 1.90049 < $\chi^2(1)$

S는 1.90049로 기각값인 3.841보다 작아 등분산 가정을 따른다는 가설을 기각할 수 없으므로 오차는 등분산 가정을 만족한다.

정규확률그림을 보면 직선에서 많이 벗어나지 않으며, w-통계량의 값이 0.98로 유의수준 0.05에서의 정규성을 따른다는 귀무가설을 기각하지 못하므로 정규성을 만족합니다.

<반응변수 변환 후 자료진단>

OBS	y	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	i1	i2	i3	lny	yhat	e1	rl	cd	tl
23	2.83	11.39	5	9	50	12	8	0.00	0.09	11.1	2	0	1	0	1.04028	1.34153	-0.30125	-1.18963	0.03791	-1.20006
24	1.81	22.00	5	15	60	12	7	0.00	0.00	6.8	2	0	1	0	0.59333	0.57860	0.01473	0.05833	0.00009	0.05715
25	9.23	3.58	23	6	40	12	2	0.56	2.51	53.0	4	0	0	1	2.22246	2.62474	-0.40228	-2.45772	0.98856	-2.76519
26	8.60	3.23	13	6	45	12	2	0.31	0.93	17.3	2	0	0	1	2.15176	2.01775	0.13401	0.53528	0.00833	0.52750

max|ti| = 2.765<

3.65=t(0.05;n=39,p'=14)

max|Di| = 0.989

반응변수 변환 후 자료진단 결과, 25번째 관측값의 절대값 t값과 cook의 D값이 가장 큽니다. 유의수준 0.05의 검정에 대한 기각값이 3.65이고, cook의 D값이 0.989이므로 이상점은 없고, 25번째 관측값이 그 하나만으로도 예외적인 결과를 낳을 수 있는 영향력을 갖고 있음을 알 수 있었습니다.

3. 모형개발

<다중공선성>

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	4.08391	1.70319	2.40	0.0243	0
x1	1	-0.01756	0.00827	-2.12	0.0437	1.70660
x2	1	0.00357	0.00841	0.42	0.6753	10.56391
x3	1	-0.02640	0.02843	-0.93	0.3619	1.93128
x4	1	-0.03772	0.02024	-1.86	0.0742	6.04126
x5	1	-0.01249	0.14818	-0.08	0.9335	1.96648
x6	1	0.00270	0.04019	0.07	0.9470	6.41795
x7	1	-0.22258	0.31789	-0.70	0.4903	7.36252
x8	1	0.16222	0.13016	1.25	0.2242	2.92901
x9	1	0.00550	0.01055	0.52	0.6068	4.16463
x10	1	-0.00404	0.07034	-0.06	0.9546	3.94795
i1	1	-0.08427	0.42830	-0.20	0.8456	9.06838
i2	1	-0.37572	0.27399	-1.37	0.1825	8.29558
i3	1	-0.22431	0.24178	-0.93	0.3624	5.74563

Collinearity Diagnostics (intercept adjusted)															
			Proportion of Variation												
Number	Eigenvalue	Condition Index	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	i1	i2	i3
1	4.50040	1.00000	0.00071831	0.00336	0.00012959	0.00332	0.00084022	0.00427	0.00420	0.00004005	0.00384	0.00788	0.00325	0.00026292	0.00451
2	2.69732	1.29168	0.00872	0.00137	0.00772	0.00671	0.00485	0.00022987	0.00128	0.00892	0.01163	0.00416	0.00015304	0.00068531	0.00002702
3	2.03129	1.48847	0.00078446	0.00198	0.00001833	0.00075496	0.01000	0.00318	0.00622	0.00212	0.00382	0.00066518	0.00833	0.02387	0.01415
4	1.07308	2.04791	0.06625	0.00048837	0.00577	0.00272	0.28543	0.00410	6.594468E-7	0.03825	0.00210	0.01054	0.00084412	0.00312	0.00044979
5	0.89599	2.24333	0.07828	0.00279	0.17222	0.02043	0.06868	0.04592	0.00317	0.01278	0.00001199	0.00436	0.00025212	0.00288	0.00751
6	0.52223	2.92559	0.46308	0.00097225	0.16535	0.03901	0.10451	0.00003999	0.00381	0.06138	0.00115	0.02034	0.00002238	0.00416	0.00456
7	0.44243	3.18936	0.01742	0.00129	0.29972	0.00954	0.03147	0.04456	0.00095679	0.04549	0.21019	0.00581	0.00015	0.00231	0.00016319
8	0.30929	3.81452	0.20957	0.00233	0.04594	0.08279	0.02488	0.04887	0.00016767	0.37095	0.15194	0.00278	0.00012896	0.01030	0.00718
9	0.23668	4.36947	0.09932	0.00022709	0.07159	0.00076951	0.00200	0.00333	0.05714	0.06475	0.00003790	0.65974	0.01227	0.00021265	0.04727
10	0.16251	5.26246	0.01611	0.06104	0.03995	0.00645	0.04339	0.02561	0.10895	0.00126	0.01679	0.00491	0.12707	0.06215	0.43750
11	0.07866	7.59316	0.00259	0.10486	0.01635	0.22400	0.09746	0.17157	0.46796	0.18468	0.17566	0.10528	0.28550	0.01389	0.19352
12	0.06023	8.64407	0.00061819	0.78465	0.01101	0.08929	0.03228	0.04912	0.25236	0.00083	0.04080	0.12943	0.24542	0.15212	0.21868
13	0.05070	9.42171	0.01079	0.03466	0.13495	0.40721	0.29443	0.59899	0.03373	0.15845	0.37194	0.03410	0.31561	0.72411	0.07748

분산팽창인자 중 10을 넘는 값이 1개 있습니다. 하지만 상태수가 30을 넘지 않고, 분산비율 역시 0.8을 넘는 값이 없으므로 다중공선성은 존재하지 않았습니다.

<변수선택>

다중공선성은 존재하지 않았지만 변수의 개수를 간소화하기 위해 변수선택을 했습니다.

1) 모든 가능한 회귀

4	0.6601	0.6201	2.3930	0.08141	x1 x4 x7 x9
4	0.6583	0.6180	2.5650	0.08186	x1 x4 x5 x9
4	0.6581	0.6179	2.5790	0.08190	x1 x4 x6 x9
5	0.7109	0.6671	-0.2971	0.07135	x1 x3 x4 x8 i2
5	0.7046	0.6599	0.2817	0.07290	x1 x4 x8 x9 i2
5	0.7029	0.6579	0.4385	0.07332	x1 x4 x7 x8 i2

수정결정계수가 가장 큰 모형은 x1, x3, x4, x8, i2입니다.

2) 전진선택법

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	3.45433	0.66035	2.11546	27.36	<.0001
x1	-0.01648	0.00684	0.44835	5.80	0.0216
x3	-0.03355	0.02308	0.16339	2.11	0.1552
x4	-0.03309	0.01057	0.75828	9.81	0.0036
x9	0.01248	0.00680	0.26012	3.36	0.0754

선택된 변수는 x1, x3, x4, x9입니다.

3) 후진제거법

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	3.70176	0.46948	4.46159	62.17	<.0001
x1	-0.02087	0.00606	0.85123	11.86	0.0015
x4	-0.03910	0.00832	1.58341	22.06	<.0001
x8	0.21660	0.08247	0.49508	6.90	0.0128
i2	-0.22369	0.09179	0.42615	5.94	0.0202

선택된 변수는 x1, x4, x8, i2

4) 단계적 회귀방법

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	3.23401	0.65287	1.95730	24.54	<.0001
x1	-0.02098	0.00620	0.91210	11.43	0.0018
x4	-0.03411	0.01071	0.80934	10.15	0.0030
x9	0.01422	0.00680	0.34863	4.37	0.0439

선택된 변수는 x1, x4, x9

<가능한 모든 회귀모형>

Number in Model	R-Square	Adjusted R-Square	C(p)	MSE	Variables in Model
4	0.7004	0.6652	-1.3281	0.07176	x1 x4 x8 i2
4	0.6773	0.6393	0.8099	0.07731	x1 x3 x4 x9
4	0.6734	0.6349	1.1681	0.07824	x1 x4 x8 x9
4	0.6700	0.6312	1.4808	0.07905	x1 x3 x4 i2
4	0.6674	0.6282	1.7227	0.07968	x1 x4 x9 i2
4	0.6667	0.6275	1.7845	0.07984	x1 x4 x8 i3
4	0.6644	0.6249	1.9989	0.08039	x1 x3 x4 x8
4	0.6638	0.6243	2.0506	0.08053	x1 x4 x9 x10
4	0.6630	0.6234	2.1251	0.08072	x1 x4 x9 i1
4	0.6629	0.6233	2.1333	0.08074	x1 x2 x4 x9
4	0.6601	0.6201	2.3930	0.08141	x1 x4 x7 x9
4	0.6583	0.6180	2.5650	0.08186	x1 x4 x5 x9
4	0.6581	0.6179	2.5790	0.08190	x1 x4 x6 x9
5	0.7109	0.6671	-0.2971	0.07135	x1 x3 x4 x8 i2
5	0.7046	0.6599	0.2817	0.07290	x1 x4 x8 x9 i2
5	0.7029	0.6579	0.4385	0.07332	x1 x4 x7 x8 i2

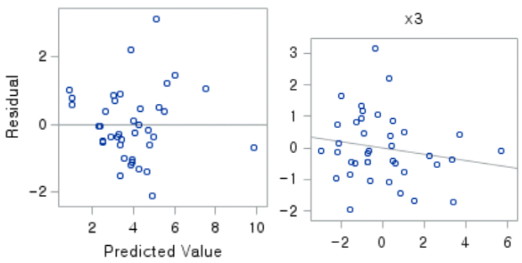
전진, 후진, 단계적 선택법에서 모두 다른 모형을 제시하였고, 이를 해결하기 위해 가능한 모든 회귀모형의 수정결정계수를 비교하였다. 수정결정계수를 기준으로 볼 때 모든 가능 한 회귀 모형 중 수정결정계수가 가장 큰 모형은 **lny=x1 x3 x4 x8 I2**으로 0.6671의 값을 갖는다. 또한 C_p <p+1이므로 고려대상에 포함되므로 이 모형을 선택하였다.

<lny= x1 x3 x4 x8 I2의 총괄분석>

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	5.78964	1.15793	16.23	<.0001
Error	33	2.35455	0.07135		
Corrected Total	38	8.14418			

Root MSE	0.26711	R-Square	0.7109
Dependent Mean	1.26034	Adj R-Sq	0.6671
Coeff Var	21.19387		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.85386	0.48832	7.89	<.0001
x1	1	-0.01782	0.00666	-2.68	0.0115
x3	1	-0.02514	0.02298	-1.09	0.2818
x4	1	-0.03820	0.00834	-4.58	<.0001
x8	1	0.18703	0.08656	2.16	0.0381
i2	1	-0.21226	0.09212	-2.30	0.0277



총괄분석 결과, 설명변수 x3에 대한 p-값이0.31으로 유의수준 0.05를 넘는다. 따라서 주어진 모형에서 다른 변수들이 모두 존재할 때 설명변수 x3의 추가설명력이 있다고 볼 수 없다. 추가변수 그림에서도 x3의 추가적인 설명력이 나타나지 않는 것으로 보인다. 따라서 x3를 제거하기로 하였다.

<x3 제거>

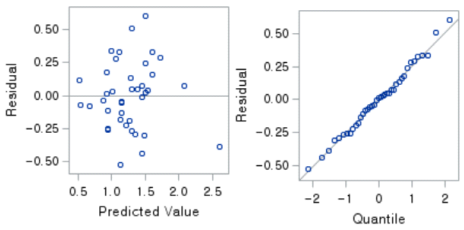
lny= x1 x4 x8 I2의 모형 선택

X3 제거로 수정결정계수는 약간 감소하였지만 모든 설명변수의 추가설명력이 유의함을 확인할 수 있다.

Root MSE	0.26789	R-Square	0.7004
Dependent Mean	1.26034	Adj R-Sq	0.6652
Coeff Var	21.25532		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.70176	0.46948	7.88	<.0001
x1	1	-0.02087	0.00606	-3.44	0.0015
x4	1	-0.03910	0.00832	-4.70	<.0001
x8	1	0.21660	0.08247	2.63	0.0128
i2	1	-0.22369	0.09179	-2.44	0.0202



정규성 검정

검정	통계량		p 값	
Shapiro-Wilk	W	0.98983	Pr < W	0.9748

정규성 검정 결과 유의수준 0.05에서 오차가 정규분포를 따른다는 가설을 기각하지 못하므로 오차는 정규성 가정을 만족한다.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	6.11029	1.52757	0.82	0.5205
Error	34	63.20919	1.85909		

u=(e**2)/(2.43998/39)
SSreg=6.11029,
S= 3.055 < 9.488= $\chi^2(4)$

스코어 검정 결과 S는 3.055로 기각값인 9.488보다 작아 등분산 가정을 따른다는 가설을 기각할 수 없으므로 오차는 등분산 가정을 만족한다.

<자료진단>

OBS	x1	x4	x8	i2	lny	e1	r1	cd	tl
1	4.99	55	0.00	0	1.52170	0.07436	0.29299	0.00196	0.28901
2	16.11	60	0.00	0	1.05082	0.03108	0.12041	0.00022	0.11865
3	9.75	60	0.00	0	1.10526	-0.04724	-0.18454	0.00065	-0.18190
4	10.65	65	0.00	0	0.82855	-0.10968	-0.44411	0.00696	-0.43880
5	20.01	70	0.00	0	0.47623	-0.07114	-0.30809	0.00657	-0.30395
6	5.97	55	1.84	1	1.92716	0.32541	1.35078	0.08632	1.36798
7	8.57	55	0.70	1	1.34807	0.04753	0.18273	0.00041	0.18011
8	5.24	55	0.38	1	1.81156	0.51082	1.99135	0.07187	2.08734
9	15.79	50	1.39	1	1.19089	-0.30389	-1.21494	0.04342	-1.22380
10	8.26	50	1.21	1	1.77156	0.15859	0.61963	0.00733	0.61393
11	7.03	60	1.85	1	1.43508	0.04877	0.21250	0.00327	0.20949
12	13.28	50	1.21	1	1.52823	0.02005	0.07869	0.00013	0.07753
13	5.40	50	0.56	1	1.56862	0.03674	0.14444	0.00045	0.14234
14	2.96	60	0.00	1	1.34807	0.27752	1.12097	0.04293	1.12535
15	11.75	55	0.60	1	0.98954	-0.22297	-0.85515	0.00813	-0.85169
16	8.86	60	0.00	1	0.68813	-0.25926	-1.01812	0.02213	-1.01868
17	9.78	60	0.10	1	0.69813	-0.25171	-0.98219	0.01787	-0.98166
18	5.49	50	0.18	1	1.43984	-0.00785	-0.03154	0.00003	-0.03107
19	8.63	55	0.00	1	1.01523	-0.13244	-0.52048	0.00587	-0.51482
20	20.31	60	0.99	1	0.93609	0.01327	0.05370	0.00010	0.05290
21	40.09	55	0.12	1	0.63658	0.11961	0.59870	0.05721	0.59297
22	11.81	60	0.00	1	0.85015	-0.03566	-0.13918	0.00036	-0.13716
23	11.39	50	0.09	1	1.04028	-0.26476	-1.05611	0.03166	-1.05796
24	22.00	60	0.00	1	0.59333	-0.07978	-0.31645	0.00259	-0.31222
25	3.58	40	2.51	0	2.22246	-0.38443	-1.93795	0.61874	-2.02431
26	3.23	45	0.93	0	2.15176	0.07528	0.30626	0.00352	0.30214
27	7.73	55	0.52	0	2.10535	0.60257	2.33658	0.08634	2.51251
28	14.41	55	0.07	0	1.07500	-0.19087	-0.73168	0.00585	-0.72658
29	11.54	45	0.09	0	2.01223	0.29117	1.18233	0.05125	1.18952
30	11.10	60	0.00	0	0.94391	-0.18042	-0.70218	0.00858	-0.69684
31	22.09	45	0.14	0	1.75267	0.24100	0.99859	0.04630	0.99855
32	9.39	55	0.00	0	1.06471	-0.29078	-1.12467	0.01860	-1.12921
33	19.49	55	0.00	0	1.08856	-0.05610	-0.21666	0.00066	-0.21359
34	21.01	55	0.10	0	0.60977	-0.52483	-2.03531	0.06567	-2.13974
35	27.16	55	0.04	0	1.32972	0.33650	1.34933	0.05605	1.36643
36	14.03	50	0.00	0	1.01523	-0.43889	-1.71235	0.05420	-1.76479
37	20.63	55	0.00	0	1.45161	0.33074	1.28150	0.02541	1.29415
38	20.06	60	0.00	0	1.11514	0.17785	0.69335	0.00872	0.68796
39	12.91	55	0.00	0	1.41585	0.13384	0.51410	0.00311	0.50846

이상점 검정 통계량 중 가장 큰 값을 가지는 27번째 관측값의 외적 스튜던트화 잔차는 2.51로 임계값인 t_{Bonf} (0.05;n=39,p'=5)=3.52보다 작으므로 이상점이 존재하지 않는다.

관측값들 중 가장 큰 cook D통계량을 가지는 25번째 관측값이 0.61874로 1보다 작은 값을 가져 영향력 관측값이 있다고 할 수 없다.

<다중공선성>

분산팽창인자의 경우 5~10을 넘지 않는다. 상태수 및 상태지표는 10 이하이며 분산비율 역시 하나의 고유값에 대하여 0.8 이상의 값을 가지는 설명변수의 수가 2개 이상이 아니다. 따라서 다중 공선성이 존재하지 않는다.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	3.70176	0.46948	7.88	<.0001	0
x4	1	-0.03910	0.00832	-4.70	<.0001	1.25489
x8	1	0.21660	0.08247	2.63	0.0128	1.44474
x1	1	-0.02087	0.00606	-3.44	0.0015	1.12638
i2	1	-0.22369	0.09179	-2.44	0.0202	1.14402

Collinearity Diagnostics (intercept adjusted)						
			Proportion of Variation			
Number	Eigenvalue	Condition Index	x4	x8	x1	i2
1	1.70944	1.00000	0.10684	0.16712	0.12487	0.06016
2	1.04351	1.27990	0.28856	0.00002950	0.00293	0.51822
3	0.78121	1.47926	0.09985	0.05860	0.84023	0.10812
4	0.46584	1.91562	0.50476	0.77425	0.03197	0.31350

4. 교호작용 추가&결론

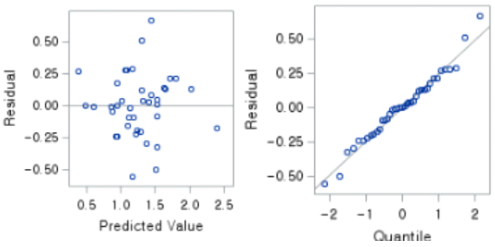
<교호작용 추가>

따라서 모형이 문제가 없는 것으로 판단하고, 이 모형에 지시변수에 대한 교호작용을 추가하여 최종모형을 결정하기로 하였다. 처음부터 교호작용을 추가한 모형으로 변수선택을 하는 것이 더 타당할 수 있겠으나, 원래의 설명변수 13개에서 3개의 지시변수와 10개의 설명변수의 교호작용을 모두 추가하면 관측수 39를 넘는 43개의 설명변수를 가진 모형을 가지기 때문에 변수 선택 이후에 교호작용을 추가하게 되었다.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	5.86345	0.83764	11.39	<.0001
Error	31	2.28073	0.07357		
Corrected Total	38	8.14418			

Root MSE	0.27124	R-Square	0.7200
Dependent Mean	1.26034	Adj R-Sq	0.6567
Coeff Var	21.52136		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.08796	0.64157	6.37	<.0001
x1	1	-0.01743	0.01074	-1.62	0.1148
x4	1	-0.04653	0.01091	-4.26	0.0002
x8	1	0.09316	0.14724	0.63	0.5316
i2	1	-0.86474	1.08630	-0.80	0.4321
x1_i2	1	-0.00709	0.01316	-0.54	0.5941
x4_i2	1	0.01199	0.01898	0.63	0.5323
x8_i2	1	0.20147	0.17989	1.12	0.2713



i2와 x1, x4, x8의 교호작용을 모두 추가한모형의 오차의 등분산성, 정규성, 모형의 선형성등은 모두 성립하는 것으로 보인다. 그러나 추가설명력이 유의하지 않은 변수들이 여럿보여 변수선택을 진행하도록 하였다.

<변수 선택>

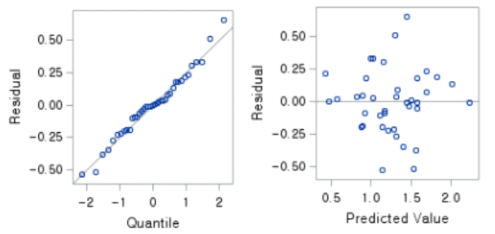
가능한 모든 모형 중 수정결정계수가 높은 모형: lny= x1 x4 I2 x8*i2 이다. (수정결정계수: 0.6747, Cp: 3.2161 < 5)

Number in Model	R-Square	Adjusted R-Square	C(p)	MSE	Variables in Model
3	0.2089	0.1411	56.5732	0.18408	i2 x4_i2 x8_i2
4	0.7090	0.6747	3.2161	0.06971	x1 x4 i2 x8_i2
4	0.7065	0.6719	3.4935	0.07031	x1 x4 x8 x1_i2
4	0.7035	0.6686	3.8207	0.07102	x1 x4 x4_i2 x8_i2

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	5.75357	1.43839	20.46	<.0001
Error	34	2.39061	0.07031		
Corrected Total	38	8.14418			

Root MSE	0.26516	R-Square	0.7065
Dependent Mean	1.26034	Adj R-Sq	0.6719
Coeff Var	21.03916		

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.15973	0.40842	10.18	<.0001
x1	1	-0.02318	0.00583	-3.98	0.0003
x4	1	-0.04600	0.00751	-6.12	<.0001
i2	1	-0.31009	0.10312	-3.01	0.0049
x8_i2	1	0.27932	0.09812	2.85	0.0074



lny= x1 x4 I2 x8*i2의 총괄분석 결과 수정결정 계수가 0.6719로 약간 감소하였지만 모든 설명변수들의 추가설명력이 유의해졌다. 또한 오차의 정규성, 등분산성, 모형의 선형성 가정이 성립함을 알 수 있다.

<정규성 검정 및 스코어검정>

정규성 검정				
검정	통계량		p 값	
Shapiro-Wilk	W	0.980193	Pr < W	0.7106

정규성 검정결과 유의수준 0.05를 넘으므로 정규성 가정을 만족한다.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	3.63537	0.90884	0.34	0.8497
Error	34	91.12878	2.68026		

u=(e7**2)/(2.37020/39)

u= x1 x4 i2 x8*I2로 적합 시

SSreg=3.63537,

S=1.818 < 9.488=$\chi^2(4)$이므로 오차가 등분산이라는 귀무가설을 기각하지 못한다. 따라서 오차는 등분산 가정을 만족한다.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.67420	0.67420	0.27	0.6097
Error	37	94.08995	2.54297		
Corrected Total	38	94.76415			

또한 스코어 검정에서 u=yhat의 경우에도 SSreg=0.67420, S=0.67420/2<3.841= $\chi^2(1)$ 이므로 오차의 등분산 가정을 만족한다.

<자료진단>

OBS	x1	x4	i2	lny	x8_i2	e7	r7	cd7	t7
1	4.99	55	0	1.52170	0.00	0.00763	0.03031	0.00002	0.02986
2	16.11	60	0	1.05082	0.00	0.02456	0.09656	0.00014	0.09514
3	9.75	60	0	1.10526	0.00	-0.06845	-0.27123	0.00139	-0.26750
4	10.65	65	0	0.82855	0.00	-0.09430	-0.38702	0.00522	-0.38213
5	20.01	70	0	0.47623	0.00	0.00039	0.00168	0.00000	0.00165
6	5.97	55	1	1.92716	1.84	0.23197	1.03055	0.07984	1.03152
7	8.57	55	1	1.34807	0.70	0.03158	0.12324	0.00019	0.12144
8	5.24	55	1	1.81156	0.38	0.50724	2.00392	0.07069	2.10227
9	15.79	50	1	1.19089	1.39	-0.38095	-1.58585	0.10467	-1.62355
10	8.26	50	1	1.77156	1.21	0.07542	0.30448	0.00252	0.30038
11	7.03	60	1	1.43508	1.85	-0.00833	-0.03798	0.00013	-0.03742
12	13.28	50	1	1.52823	1.21	-0.05152	-0.20870	0.00125	-0.20574
13	5.40	50	1	1.56862	0.56	-0.01227	-0.04865	0.00005	-0.04793
14	2.96	60	1	1.34807	0.00	0.32703	1.35748	0.07413	1.37515
15	11.75	55	1	0.98954	0.60	-0.22530	-0.87673	0.00855	-0.87367
16	8.86	60	1	0.68813	0.00	-0.19612	-0.79302	0.01758	-0.78860
17	9.78	60	1	0.69813	0.10	-0.19272	-0.77134	0.01388	-0.76665
18	5.49	50	1	1.43984	0.18	-0.03282	-0.13235	0.00047	-0.13042
19	8.63	55	1	1.01523	0.00	-0.10436	-0.41853	0.00425	-0.41340
20	20.31	60	1	0.93609	0.99	0.04077	0.16538	0.00080	0.16300
21	40.09	55	1	0.63658	0.12	0.21284	1.09559	0.20335	1.09893
22	11.81	60	1	0.85015	0.00	0.03429	0.13788	0.00048	0.13588
23	11.39	50	1	1.04028	0.09	-0.27046	-1.08888	0.03082	-1.09195
24	22.00	60	1	0.59333	0.00	0.01371	0.05614	0.00011	0.05532
25	3.58	40	0	2.22246	0.00	-0.01429	-0.06221	0.00025	-0.06129
26	3.23	45	0	2.15176	0.00	0.13690	0.56537	0.01208	0.55963
27	7.73	55	0	2.10535	0.00	0.65481	2.57106	0.09879	2.82212
28	14.41	55	0	1.07500	0.00	-0.22067	-0.85754	0.00776	-0.85412
29	11.54	45	0	2.01223	0.00	0.19003	0.76910	0.01679	0.76438
30	11.10	60	0	0.94391	0.00	-0.19851	-0.78381	0.01067	-0.77927
31	22.09	45	0	1.75267	0.00	0.17506	0.73095	0.02302	0.72585
32	9.39	55	0	1.06471	0.00	-0.34734	-1.35735	0.02379	-1.37502
33	19.49	55	0	1.08856	0.00	-0.08933	-0.34981	0.00169	-0.34525
34	21.01	55	0	0.60977	0.00	-0.53289	-2.09697	0.06990	-2.21403
35	27.16	55	0	1.32972	0.00	0.32965	1.34146	0.05557	1.35801
36	14.03	50	0	1.01523	0.00	-0.51925	-2.03744	0.06086	-2.14229
37	20.63	55	0	1.45161	0.00	0.30015	1.17951	0.02131	1.18657
38	20.06	60	0	1.11514	0.00	0.18046	0.71367	0.00920	0.70842
39	12.91	55	0	1.41585	0.00	0.08541	0.33199	0.00118	0.32760

자료진단 결과 가장 큰 이상점 통계량 값은 2.822로 기각역인 3.52를 넘지 못해 이상점이 없다는 것을 확인하였다. 영향력 관측값에서도 가장 큰 값인 0.203이 1보다 매우 작은 값이므로 영향력 관측값은 없다.

<다중공선성>

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	4.15973	0.40842	10.18	<.0001	0
x1	1	-0.02318	0.00583	-3.98	0.0003	1.07213
x4	1	-0.04600	0.00751	-6.12	<.0001	1.05196
i2	1	-0.31009	0.10312	-3.01	0.0049	1.48623
x8_i2	1	0.27932	0.09812	2.85	0.0074	1.49076

Collinearity Diagnostics (intercept adjusted)						
Number	Eigenvalue	Condition Index	Proportion of Variation			
			x1	x4	i2	x8_i2
1	1.65357	1.00000	0.07370	0.01001	0.16832	0.17779
2	1.13268	1.20825	0.23836	0.50010	0.05377	0.01429
3	0.78688	1.44963	0.68670	0.44019	0.00428	0.04372
4	0.42687	1.96817	0.00125	0.04969	0.77362	0.76419

분산팽창인자의 경우 5~10을 넘지 않는다. 상태수 및 상태지표는 10 이하이며 분산비율 역시 0.8 이상의 값을 가지는 설명변수가 존재하지 않았다. 따라서 다중 공선성이 존재하지 않는다.

결론

최종모형: $\ln y = 4.15973 - 0.02318x_1 - 0.04600x_4 - 0.31009i_2 + 0.27932x_8 * i_2$ 주고속도인 경우(I2=1): $\ln y = 3.84964 - 0.02318x_1 - 0.04600x_4 + 0.27932x_8$ 주고속도로가 아닌 도로의 경우(I2=0): $\ln y = 4.15973 - 0.02318x_1 - 0.04600x_4$
Y: 100만 마일의 자동차 주행거리당 사고의 횟수 X1: 구간의 길이 X4: 제한속도 X8: 1마일당 신호등이 있는 교차로의 수 I2: 1: 주고속도로 0: 기타
주고속도로일 경우 다른 변수들이 모두 고정되어있을 때 구간의 길이가 한 단위 증가할 때마다 log(사고횟수)가 0.02318씩 감소한다. 또한 나머지 변수가 고정되어있을때 제한속도가 한 단위 증가할 때 log(사고횟수)는 0.406만큼 감소한다. 나머지 변수가 모두 고정되었을 때 1마일당 신호등이 있는 교차로의 수가 한 단위 증가할 때 log(사고횟수)는 0.27932 만큼 증가한다. 주고속도로가 아닌 도로의 경우 나머지 변수가 고정되어있을 때 구간의 길이가 한 단위 증가할 때 log(사고횟수)는 0.02318씩 증가하며 제한속도가 한단위 증가할 때마다 0.04600만큼 감소한다. 따라서 사고횟수는 주고속도로 일 때 구간의 길이, 제한속도, 1마일당 신호등이 있는 교차로의 수의 영향을 받으며 주고속도가 아닐 때는 1마일당 신호등이 있는 교차로의 수의 영향은 받지 않는다. 즉 **교통사고의 원인**은 주고소로에서는 구간의 길이, 제한속도, 1마일당 신호등이 있는 교차로의 수이며, 주고속도로가 아닌 경우에는 구간의 길이, 제한속도라고 할 수 있다.