Chapter 7 변수선택

지금까지 다중회귀에서 변수 선택은 다음 방법에 의존하였다.

○경험이나 이론에 의해 종속변수에 영향을 미칠 것 같은 설명변수를 선택하고 선형 회귀모형을 설정한다.

○데이터를 수집 후 회귀모형의 회귀계수를 추정하고(OLS) F-검정에 의해 모형의 유의성을 검정한다. ($H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$, 모든 설명변수는 유의하지 않다)

○F-검정(분산분석) 결과 귀무가설이 기각되면 t-검정을 이용하여 회귀계수(설명변수)의 유의성 검정을 한다.

○유의하지 않은 설명변수를 하나씩 제외하면서 모든 변수가 유의할 때까지 계속한다. 어떤 설명변수를 제외할 것인가? 가장 유의하지 않은 설명변수, 즉 t-검정통계량 값이 가장 작은 것(p-value가 가장 큰 것)을 제외한다.

○비록 유의확률이 다른 것에 비해 다소 적더라도 분석자의 판단에 의해 설명변수를 제외할 수 있다.

EXAMPLE 예제 자료

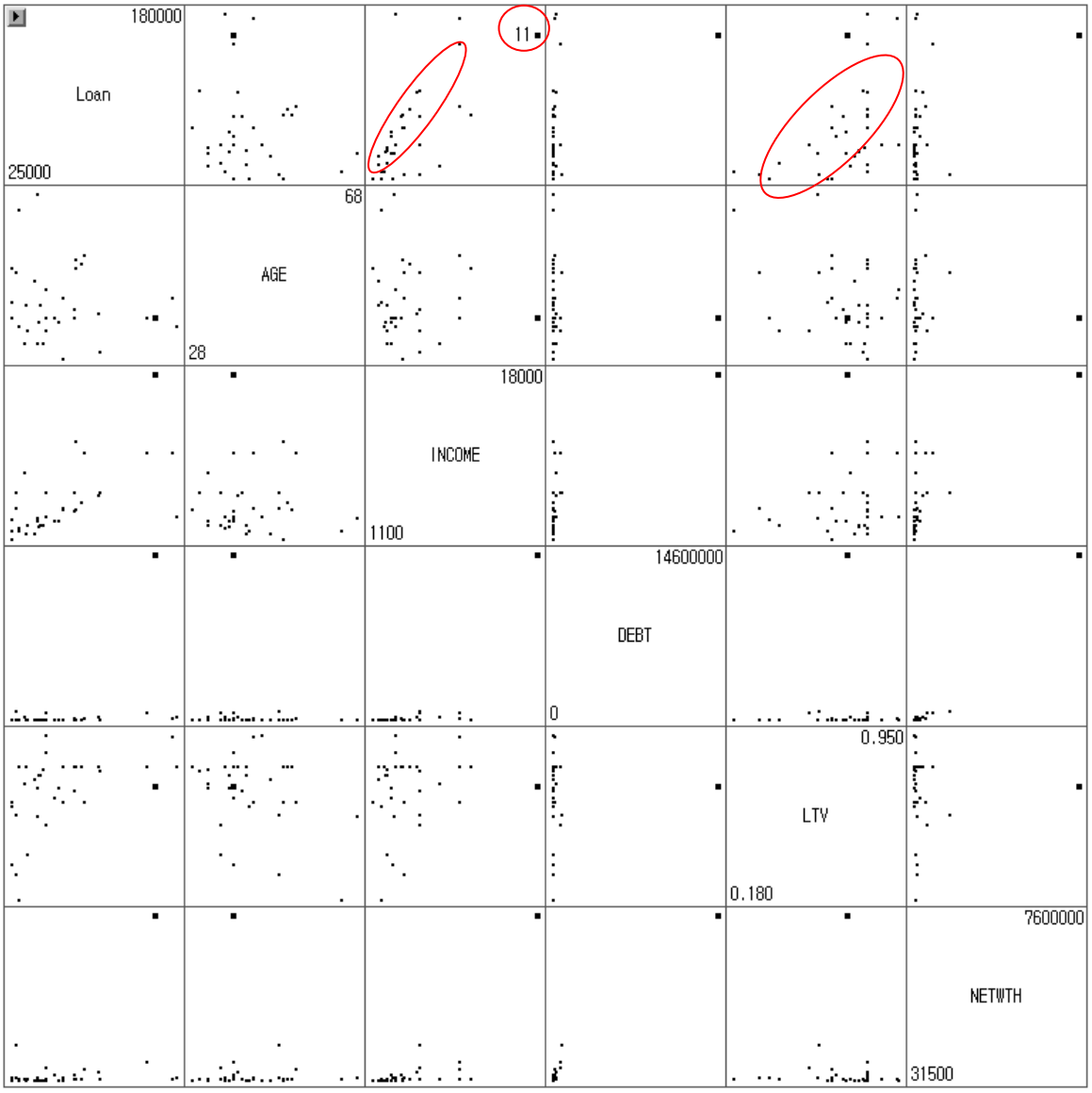
변수 8개, 설명 변수 7개(지시 변수 2개 포함), 본 장에서는 분석의 간편함을 위하여 지시변수(MARRIAGE, FIXED)를 제외하고 분석하기로 한다. **LOAN.txt**

```
1.Loan: Loan amount.(Y)
2.AGE
3.INCOME
4.MARRIAGE: =1 if married and 0 otherwise.
5 DEBT: debt
6 LTV: loan to value ratio.
7 NETWTH: lendee's net worth.
8 FIXED: =1 if the loan assumes a fixed interest rate
         0 if a variable interest rate assumption is used.

  40000  35   2600  1     17300  0.395   154000  0
  51200  37   2400  1      4000  0.260    64000  0
```

```
DATA LOAN;
    INPUT Loan AGE INCOME MARRIAGE DEBT LTV NETWTH FIXED;
    CARDS;
    40000  35   2600  1     17300  0.395   154000  0
```

산점도 행렬을 그려보자.



종속변수 LOAN에 영향을 미칠 것 같은 설명변수는 INCOME, LTV이고 이상치가 존재하는 것 같다. 설명변수 INCOME과 LTV는 매우 약한 상관 관계가 존재하는 것 같다. 다중공성선성 문제가 발생하지 않을까 의심된다.

이제 모형이 유의한지(F-검정) 각 설명변수가 유의한지 검정해 보자. 변수 선택을 먼저 해야 하나? 아니면 다중공성선 문제 진단을 먼저 해야 하나? 일반적으로 결과는 같으므로 (likely) 유의한 설명변수를 먼저 선택하고 유의한 설명변수만으로 다중공선성 문제 진단하면 더 간편하다.

7.1 수작업 하기

$$Loan = \beta_0 + \beta_1 * age + \beta_2 * income + \beta_3 * debt + \beta_4 * ltv + \beta_5 * netwth + e$$

7.1.1 모형 유의성 검정(F-검정)

```
proc reg data=loan;
    model Loan=AGE INCOME DEBT LTV NETWTH;
run;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	32723284029	6544656806	6.88	0.0002
Error	29	27570360931	950702101		
Corrected Total	34	60293644960			

유의확률이 0.0002이므로 귀무가설 $H_0 : \beta_1 = \beta_2 = ... = \beta_5 = 0$ 이 기각된다. 그러므로 고려한 설명변수 5개 중 적어도 하나 이상은 유의하다.

7.1.2 설명변수 유의성 검정(t-검정)

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-19733	41118	-0.48	0.6349
AGE	1	-77.23684	643.19795	-0.12	0.9052
INCOME	1	7.40606	2.31683	3.20	0.0033
DEBT	1	0.00306	0.01282	0.24	0.8130
LTV	1	85866	35225	2.44	0.0211
NETWTH	1	-0.00727	0.02557	-0.28	0.7781

각 설명변수의 유의성을 t-검정 하면 (AGE, INCOME, NETWTH)은 유의하지 않다. 그러나 모형에서 동시에 제외해서는 안된다. 가장 유의하지 않은 설명변수를(유의확률이 가장 크거나 t-값이 가장 작은 것) 제외하고 다시 설명변수 유의성을 검정한다.

우선 유의확률이 가장 큰 AGE 설명변수를 제외한다. 그러나 AGE 설명변수가 DEBT에 비해 해석이 용이하거나 더 중요하다고 판단되면 분석자가 DEBT를 먼저 제외할 수 있다.

AGE 제외

```
proc reg data=loan;
    model Loan=INCOME DEBT LTV NETWTH;
run;
```

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-23778	23183	-1.03	0.3132
INCOME	1	7.38307	2.27066	3.25	0.0028
DEBT	1	0.00326	0.01249	0.26	0.7957
LTV	1	87239	32766	2.66	0.0123
NETWTH	1	-0.00758	0.02501	-0.30	0.7638

DEBT 제외

```
proc reg data=loan;
    model Loan=INCOME LTV NETWTH;
run;
```

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-25752	21586	-1.19	0.2419
INCOME	1	7.27372	2.19796	3.31	0.0024
LTV	1	89406	31220	2.86	0.0074
NETWTH	1	-0.00123	0.00583	-0.21	0.8336

NETWTH 제외

```
proc reg data=loan;
    model Loan=INCOME LTV;
run;
```

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-25646	21256	-1.21	0.2365
INCOME	1	6.93401	1.48022	4.68	<.0001
LTV	1	90854	30003	3.03	0.0048

설명변수 INCOME, LTV은 유의하다. 이제 다중공선성 문제와 이상치(영향치) 진단을 거친 후 잔차분석을 한다. 다중회귀분석 절차는 마지막에 다시 논의하기로 한다.

tentative 추정 모형: $LOAN = -25646 + 6.934 * INCOME + 90854 * LTV$

INCOME과 LTV 중 어떤 설명변수가 종속변수 LOAN에 더 큰 영향을 미치는가? 이에 대한 대답은 표준화 회귀계수를 구하면 된다. 설명변수 INCOME의 영향력이 더 크다.

```
proc reg data=loan;
    model Loan=INCOME LTV/stb;
run;
```

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	-25646	21256	-1.21	0.2365	0
INCOME	1	6.93401	1.48022	4.68	<.0001	0.57110
LTV	1	90854	30003	3.03	0.0048	0.36917

7.1.3 다중공선성 문제 진단 먼저

```
proc reg data=loan;
    model Loan=AGE INCOME DEBT LTV NETWTH/vif collin;
run;
```

분산팽창지수에 의하면 DEBT와 NETWTH가 다중공선성 문제를 일으키는 것으로 판단된다. 이는 산점도 행렬에서 INCOME과 LTV가 문제가 있을 것으로 판단되었는데... 아직 발견되지 않고 있다. DEBT와 NETWTH가 어떤 변수와 다중공선성 문제를 일으키는지 알아보기 위하여 상태지수를 살펴보자.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-19733	41118	-0.48	0.6349	0
AGE	1	-77.23684	643.19795	-0.12	0.9052	1.17243
INCOME	1	7.40606	2.31683	3.20	0.0033	2.30926
DEBT	1	0.00306	0.01282	0.24	0.8130	35.36941
LTV	1	85866	35225	2.44	0.0211	1.29927
NETWTH	1	-0.00727	0.02557	-0.28	0.7781	38.61155

상태지수가 10이상인 경우는 마지막 두 개 행이며 문제가 되는 변수는 (DEBT, NETWTH)와 (AGE, LTV)이다. 다중공선성 문제 해결 방법으로 문제가 되는 설명변수를 제거하는 방법을 사용하자.

Collinearity Diagnostics

Condition Index	Proportion of Variation: Intercept	AGE	INCOME	DEBT	LTV	NETWTH
1.00000	0.00080020	0.00172	0.00757	0.00049822	0.00210	0.00057776
1.57667	0.00080262	0.00190	0.00028116	0.00656	0.00212	0.00474
5.68621	0.00797	0.03180	0.83710	0.01070	0.00107	0.00470
7.96028	0.00000493	0.20358	0.06073	0.01053	0.37655	0.00354
18.09124	0.09952	0.23137	0.05428	0.82943	0.00178	0.82024
20.02759	0.89090	0.52964	0.04003	0.14229	0.61638	0.16620

종속변수 LOAN에 상관관계가 높은(직선적 설명력이 높은 변수) AGE 변수보다 상관 관계가 높은 LTV, NETWTH 변수보다 상관 관계가 높은 BEBT를 선택하면 된다. 물론 둘 다 유의한 경우 주관적으로 AGE나 DEBT를 선택할 수도 있다.

```
proc corr;
    var loan;
    with AGE DEBT LTV NETWTH;
run;
```

	Loan
AGE	-0.16143 0.3542
DEBT	0.38759 0.0214
LTV	0.47529 0.0039
NETWTH	0.38315 0.0231

```
proc reg data=loan;
    model Loan=INCOME LTV DEBT/VIF COLLIN;
run;
```

더 이상 다중공선성 문제는 없다.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-25967	21706	-1.20	0.2406	0
INCOME	1	7.14292	2.09665	3.41	0.0018	2.01444
LTV	1	90180	30836	2.92	0.0064	1.06055
DEBT	1	-0.00041633	0.00291	-0.14	0.8873	1.94559

Collinearity Diagnostics

Eigenvalue	Condition Index	Proportion of Variation: Intercept	INCOME	LTV	DEBT
2.91464	1.00000	0.00562	0.01749	0.00550	0.01415
0.94470	1.75643	0.00552	0.00413	0.00544	0.40531
0.11225	5.09565	0.05274	0.97684	0.04306	0.57939
0.02841	10.12805	0.93611	0.00154	0.94600	0.00116

유의하지 않은 설명변수 DEBT를 제외하면 7.1.2절의(변수선택 먼저) 결과와 동일하다. 그러므로 분석의 간편화를 위하여 유의한 변수를 먼저 선택하고 다중공선성 진단을 하면 된다. 그러나 돌다리도 두드려 가는 심정으로 다중공선성 먼저, 나중 변수 선택 수순을 밟아가는 것도 괜찮다.

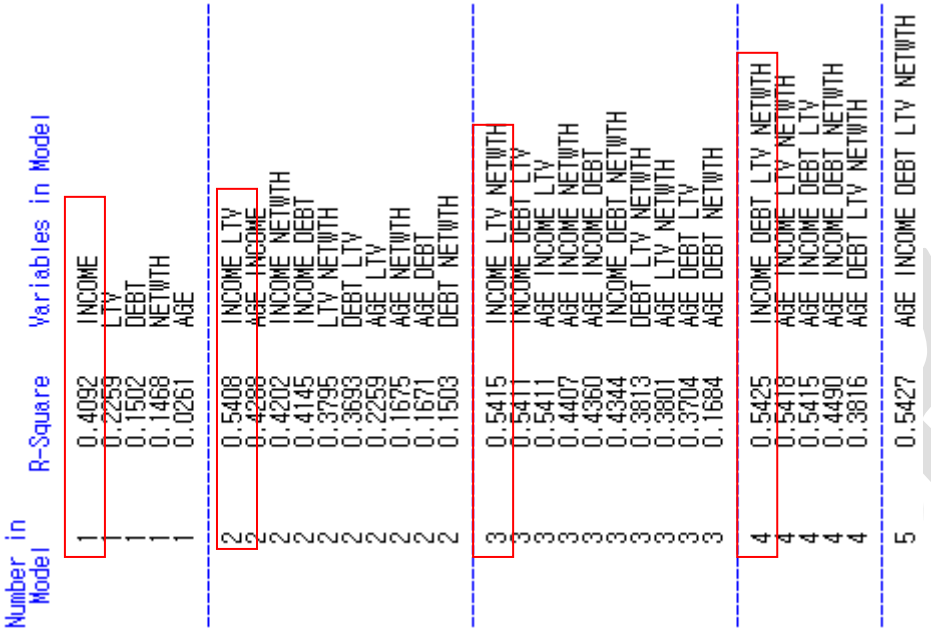
7.2 통계량 이용하기

7.2.1 결정 계수($R_p^2 = \frac{SSR_p}{SST} = 1 - \frac{SSE_p}{SST}$)

R_p^2 는 설명 변수들의 설명력의 정도를 나타내는 수치이므로 변수 선택의 지표가 된다. 설명 변수의 수가 같은 경우 어떤 변수 그룹이 설명력이 높은가를 쉽게 알아보는 사용할 수 있으나 검정 통계량은 존재하지 않는 단점이 있다. 또한 R_p^2 는 설명변수의 수(p)가 증가할 때 마다 항상 증가하므로 변수의 개수가 다른 경우에는 수정 결정 계수를 사용하는 것이 좋다.

```
PROC REG DATA=LOAN;
    MODEL Loan=AGE INCOME DEBT LTV NETWTH/SELECTION=RSQUARE;
RUN;
```

혹은 Residual Mean Square ($RMS = \frac{SSE_p}{n-p-1}$)을 최소화하는 모형이 좋은 모형이다. 이것은 결정계수를 이용한 결과와 동일하므로 자주 이용되지는 않는다.



Number in Model	R-Square	Variables in Model
1	0.4092	INCOME
1	0.2259	LTV
1	0.1502	DEBT
1	0.1468	NETWTH
1	0.0261	AGE
2	0.5408	INCOME LTV
2	0.4288	AGE INCOME
2	0.4202	INCOME NETWTH
2	0.4145	INCOME DEBT
2	0.3795	LTV NETWTH
2	0.3693	DEBT LTV
2	0.2259	AGE LTV
2	0.1675	AGE NETWTH
2	0.1671	AGE DEBT
2	0.1503	DEBT NETWTH
3	0.5415	INCOME LTV NETWTH
3	0.5411	INCOME DEBT LTV
3	0.5411	AGE INCOME LTV
3	0.4407	AGE INCOME NETWTH
3	0.4360	AGE INCOME DEBT
3	0.4344	INCOME DEBT NETWTH
3	0.3813	DEBT LTV NETWTH
3	0.3801	AGE LTV NETWTH
3	0.3704	AGE DEBT LTV
3	0.1684	AGE DEBT NETWTH
4	0.5425	INCOME DEBT LTV NETWTH
4	0.5418	AGE INCOME LTV NETWTH
4	0.5415	AGE INCOME DEBT LTV
4	0.4490	AGE INCOME DEBT NETWTH
4	0.3816	AGE DEBT LTV NETWTH
5	0.5427	AGE INCOME DEBT LTV NETWTH

설명 변수 개수에 따라 설명력이 가장 큰 변수 그룹을 알 수 있다. 각 설명 변수의 유의성을 검정한 것은 아니다. 이것을 이용하면 변수의 개수에 따른 가장 좋은 변수 결합 조건을 얻을 수 있다. 가장 유의하지 않은 설명변수를 제외하는 수작업의 결과와 각 단계의 결정계수가 가장 높은 곳과 일치한다.

7.2.2 수정 결정계수 이용 $R^2_{adj} = 1 - \frac{SSE_p/(n-p-1)}{SST/(n-1)}$

수정(adjusted) 결정계수 R^2_{adj} 는 R^2_p 의 문제점(유의하지 않은 설명변수가 삽입되어도 항상 증가)을 해결하였으므로 R^2_{adj} 값이 가장 큰 설명 변수 그룹을 선택하면 된다. 그러나 설명변수의 증가로 줄어든 SSE의 값과 $(n-p-1)$ 의 감소가 상쇄되므로 설명력의 지표로 좋은 것은 아니다. 그리고 유의성을 검정할 검정통계량이 존재하지 않는다. 수정결정계수는 설명변수가 서로 다른 회귀모형의 설명력을 비교할 때 사용된다. 그러나 여전히 차이에 대한 검정 방법은 존재하지 않는다.

```
MODEL Loan=AGE INCOME DEBT LTV NETWTH/SELECTION=ADJSQ;
```

Number in Model	Adjusted R-Square	R-Square	Variables in Model
2	0.5121	0.5408	INCOME LTV
3	0.4971	0.5415	INCOME LTV NETWTH
3	0.4967	0.5411	INCOME DEBT LTV
3	0.4966	0.5411	AGE INCOME LTV
4	0.4815	0.5425	INCOME DEBT LTV NETWTH
4	0.4807	0.5418	AGE INCOME LTV NETWTH
4	0.4803	0.5415	AGE INCOME DEBT LTV
5	0.4639	0.5427	AGE INCOME DEBT LTV NETWTH
2	0.3931	0.4288	AGE INCOME
1	0.3913	0.4092	INCOME
3	0.3866	0.4407	AGE INCOME NETWTH
2	0.3839	0.4202	INCOME NETWTH
3	0.3814	0.4360	AGE INCOME DEBT
3	0.3797	0.4344	INCOME DEBT NETWTH
2	0.3779	0.4145	INCOME DEBT
4	0.3756	0.4490	AGE INCOME DEBT NETWTH
2	0.3407	0.3795	LTV NETWTH
2	0.3299	0.3693	DEBT LTV
3	0.3214	0.3813	DEBT LTV NETWTH
3	0.3201	0.3801	AGE LTV NETWTH
3	0.3095	0.3704	AGE DEBT LTV
4	0.2992	0.3816	AGE DEBT LTV NETWTH
1	0.2024	0.2259	LTV
2	0.1775	0.2259	AGE LTV
1	0.1245	0.1502	DEBT
1	0.1209	0.1468	NETWTH
2	0.1155	0.1675	AGE NETWTH
2	0.1151	0.1671	AGE DEBT
2	0.0972	0.1503	DEBT NETWTH
3	0.0879	0.1684	AGE DEBT NETWTH
1	-.0035	0.0261	AGE

어느 변수 군을 택할 것인가? 설명력이 비슷한 그룹 중(빨간 박스 안) 분석자가 선택하면 된다. 아직 각 변수의 유의성을 검정한 것은 아니다.

7.2.3 Mallow C_p 이용 $C_p = \frac{SSE_p}{MSE(Full)} - n + 2(p+1)$

C_p 값이 p (설명변수 개수)와 근사한 경우 좋은 회귀 모형으로 판단한다. 여전히 이 방법에서도 각 설명변수의 유의성에 대한 검정을 실시한 것은 아니다. $MSE(F) = \sigma^2$ 이고 만약 모형이 적합하다면 $\frac{SSE_p}{n-p-1} \cong \sigma^2$ 이므로 $C_p \cong (p+1)$ 일 것이다.

```
PROC REG DATA=LOAN;
     MODEL Loan=AGE INCOME DEBT LTV NETWTH/SELECTION=CP;
RUN;
```

.

Number in Model	C(p)	R-Square	Variables in Model
2	0.1225	0.5408	INCOME LTV
3	2.0804	0.5415	INCOME LTV NETWTH
3	2.1033	0.5411	INCOME DEBT LTV
3	2.1061	0.5411	AGE INCOME LTV
4	4.0144	0.5425	INCOME DEBT LTV NETWTH
4	4.0570	0.5418	AGE INCOME LTV NETWTH
4	4.0809	0.5415	AGE INCOME DEBT LTV
5	6.0000	0.5427	AGE INCOME DEBT LTV NETWTH
1	6.4676	0.4092	INCOME
2	7.2227	0.4288	AGE INCOME
2	7.7737	0.4202	INCOME NETWTH
2	8.1329	0.4145	INCOME DEBT
3	8.4685	0.4407	AGE INCOME NETWTH
3	8.7715	0.4360	AGE INCOME DEBT
3	8.8702	0.4344	INCOME DEBT NETWTH
4	9.9422	0.4490	AGE INCOME DEBT NETWTH
2	10.3538	0.3795	LTV NETWTH
2	10.9997	0.3693	DEBT LTV
3	12.2394	0.3813	DEBT LTV NETWTH
3	12.3163	0.3801	AGE LTV NETWTH
3	12.9301	0.3704	AGE DEBT LTV
4	14.2185	0.3816	AGE DEBT LTV NETWTH
1	18.0933	0.2259	LTV
2	20.0929	0.2259	AGE LTV
1	22.8930	0.1502	DEBT
1	23.1098	0.1468	NETWTH
2	23.7947	0.1675	AGE NETWTH
2	23.8208	0.1671	AGE DEBT
2	24.8874	0.1503	DEBT NETWTH
3	25.7389	0.1684	AGE DEBT NETWTH
1	30.7675	0.0261	AGE

Mallow C_p 면에서는 위에서 빨간 줄을 친 변수 그룹들을 적용한 모형이 적절하다. 그러나 여전히 각 변수의 유의성은 검정되지 않았다.

7.2.4 $PRESS_p$ (Prediction REsidual Sum of Squares 예측잔차 자승합) $PRESS_p = \sum_{i=1}^{n}(Y_i - \hat{Y}_{(i)})^2$

$\hat{Y}_{(i)}$ 는 i-번째 관측치를 제외한 후 추정한 모형에 의해 추정된 예측치이므로 $(Y_i - \hat{Y}_{(i)})$ 는 제외 잔차라 할 수 있다. 예측 잔차 자승합 $PRESS_p$ 가 적을수록 좋은 회귀 모형이다.

```
PROC REG DATA=LOAN;
     MODEL Loan=AGE INCOME DEBT LTV NETWTH/P;
RUN;
```

Obs	Dep Var Loan	Predicted Value	Residual
1	40000	29670	10330
2	51200	59989	-8789
3	180000	71376	108624
4	60000	64225	-4225

Sum of Residuals	0
Sum of Squared Residuals	27570360931
Predicted Residual SS (PRESS)	3.750717E11

위의 결과는 (AGE INCOME DEBT LTV NETWTH)를 설명 변수로 사용하였을 때 PRESS

값을 구한 것이다. 이전과 달리 각 변수 군에 따라 PRESS 값들이 출력되는 옵션은 없다. 이는 PRESS에 의해 적절한 변수 군을 선택하는 것은 의미가 없는 (값의 차이가 거의 없으므로) 경우가 많다. PRESS를 주로는 사용하는 경우는 변수를 선택하는데 사용하기보다는 전혀 다른 변수 군들 중 어느 것이 더 적합한지 알아볼 때이다. (앞의 수정 결정계수 사용하는 분야도 동일하다.)

7.2.5 Comment

앞에서 살펴본 방법 (수정)결정계수, Mallow C_p, PRESS 통계량 이용 방법은 어떤 변수 군들이 좋은지에 대한 지표만을 제공할 뿐 변수의 유의성은 검정되지 않았다. 일반적으로 이 방법들을 이용할 때는 수정 결정 계수와 Mallow C_p에 의해 좋다고 간주되는 변수 군들을 몇 개 선택하고 각 변수 군에 대해 PRESS 값을 계산하여 최적 변수 군을 선택하면 된다. 이 방법들은 변수의 유의성이 검정하지 않으며 검정 통계량이 없으므로 변수 군 선택 판단에 참고 자료로 이용될 뿐이다.

PRESS는 전혀 다른 변수 군을 비교할 때(변수들이 모두 유의한 경우) 사용되므로 널리 사용된다. 또한 앞에서 언급하였듯 AIC(Akaike Information Criteria)= $n\ln(SSE/n)+2(p-1)$, SBC(Schwarz's Bayesian criterion)= $n\ln(SSE/n)+(p-1)\ln(n)$도 서로 다른 변수 군의 적합정도 비교 시 사용된다. AIC, SBC 모두 작을수록 적합도가 높다.

7.3 유의성 검정에 의한 변수 선택 방법

설명변수의 유의성을 검정하면서 변수를 선택하는 방법을 알아보기로 하자.

7.3.1 후진제거

모든 설명변수를 고려한 모형에서 유의하지 않은 설명변수를 하나씩 제거하는 방법이다.

⑴고려된 설명변수를 모두 삽입한 후 설명변수 중 가장 유의하지 않은 설명변수를 제외한다.

⑵가장 유의하지 않다는 것은 다음 검정 결과 유의하지 않고 F값이 가장 작은 설명변수를 의미한다.

$$F=\frac{MSR(X_k \mid X_1,X_2,X_{k-1},X_{K+1},\ldots X_p)}{MSE(X_1,X_2,\ldots X_p)}=\frac{MSR(X_1,X_2,\ldots X_p)-MSR(X_1,X_2,X_{k-1},X_{K+1},\ldots X_p)}{MSE(X_1,X_2,\ldots X_p)}$$

⑶모든 설명변수가 유의할 때까지 (2)과정을 반복한다.

```
PROC REG DATA=LOAN;
     MODEL Loan=AGE INCOME DEBT LTV NETWTH/SELECTION=BACKWARD;
RUN;
```

후진제거의 경우 유의수준을 지정하는 옵션이 SLS(Significant Level for Stay)이 있는데 default는 0.1(유의수준 10%)이다. `SELECTION=BACKWARD SLS=0.05;` 이것은 유의수준을 0.05로 하기 원할 때 하면 된다.

The REG Procedure
Model: MODEL1
Dependent Variable: Loan

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.5427 and C(p) = 6.0000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	32723284029	6544656806	6.88	0.0002
Error	29	27570360931	950702101		
Corrected Total	34	60293644960			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-19733	41118	218952556	0.23	0.6349
AGE	-77.23684	643.19795	13708949	0.01	0.9052
INCOME	7.40606	2.31683	9714769280	10.22	0.0033
DEBT	0.00306	0.01282	54158886	0.06	0.8130
LTV	85866	35225	5649233296	5.94	0.0211
NETWTH	-0.00727	0.02557	76899750	0.08	0.7781

Bounds on condition number: 38.612, 393.81

Backward Elimination: Step 1

Variable AGE Removed: R-Square = 0.5425 and C(p) = 4.0144

STEP3에서 최종적으로 유의수준 0.05에서 유의한 설명변수, INCOME, LTV(앞에서 수작업과 일치, 후진 제거에서 0.1이 사용되었음에도 유의한 설명변수가 동일한 것은 0.05~0.1 사이의 유의확률을 갖는 설명변수가 없었기 때문)만 남게 된다. 그리고 제외된 설명변수 순서대로 제거될 때 유의확률과 함께 출력된다.

Backward Elimination: Step 3

Variable NETWTH Removed: R-Square = 0.5408 and C(p) = 0.1225

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	32606800504	16303400252	18.84	<.0001
Error	32	27686844456	865213889		
Corrected Total	34	60293644960			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-25646	21256	1259479264	1.46	0.2365
INCOME	6.93401	1.48022	18986240118	21.94	<.0001
LTV	90854	30003	7933634585	9.17	0.0048

Bounds on condition number: 1.0358, 4.143

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

Summary of Backward Elimination

Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	AGE	4	0.0002	0.5425	4.0144	0.01	0.9052
2	DEBT	3	0.0010	0.5415	2.0804	0.07	0.7957
3	NETWTH	2	0.0007	0.5408	0.1225	0.04	0.8336

7.3.2 전진삽입

⑴고려된 설명변수 중 설명력($SSR(X_k)$)이 가장 높고 설명력이 유의하면 변수를 선택한다.

⑵이미 선택된 설명변수(X_k)의 설명 부분을 제외한 $SSR(X_l \mid X_k)$이 가장 크고 그 설명력이 유의한 경우 ($F = \frac{MSR(X_k)}{MSE(X_k)}$의 유의성 검정) X_l을 선택한다.

⑶이미 선택된 설명변수(X_k, X_l)의 설명 부분을 제외한 $SSR(X_m \mid X_k, X_l)$이 가장 크고 그 설명력이 유의한 경우 ($F = \frac{SSR(X_k, X_l) - SSR(X_k)/1}{MSE(X_k, X_l)}$ 유의성 검정) X_m을 선택한다.

⑷유의한 설명변수가 없을 때까지 [3]을 반복한다.

```
PROC REG DATA=LOAN;
     MODEL Loan=AGE INCOME DEBT LTV NETWTH/SELECTION=FORWARD;
RUN;
```

전진삽입의 경우 유의수준을 지정하는 옵션이 있다. SLE(Significant Level for Entry)이다. default는 0.5이다. 다소 큰 이유는 가능하면 많은 설명변수를 선택하기 위함이다.

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-25646	21256	1259479264	1.46	0.2365
INCOME	6.93401	1.48022	18986240118	21.94	<.0001
LTV	90854	30003	7933634585	9.17	0.0048

Bounds on condition number: 1.0358, 4.143

중간 생략. No other variable met the 0.5000 significance level for entry into the model.

결과는 이전과 동일하다. 이는 유의수준 0.05~0.5 사이에서 유의한 설명변수가 존재하지 않았기 때문이다.

7.3.3 단계삽입

단계 삽입(stepwise)은 Forward 방법과 유사하지만 한 번 선택된 설명 변수에 대해서는 유의성 검정을 다시 실시한다는 점이 다르다.

⑴고려된 설명변수 중 설명력($SSR(X_k)$)이 가장 높고 설명력이 유의하면($F = \frac{MSR(X_k)}{MSE(X_k)}$ 의 유의성 검정) 변수를 선택한다.

⑵이미 선택된 설명변수(X_k)의 설명 부분을 제외한 $SSR(X_l \mid X_k)$ 이 가장 크고 그 설명력이 유의한 ($F = \frac{SSR(X_k, X_l) - SSR(X_k)/1}{MSE(X_k, X_l)}$ 유의성 검정) 경우 X_l 을 선택한다.

⑶새로 선택된 변수의 설명 부분을 제외한 부분에 대해 이미 존재한 설명 변수의 유의성 $SSR(X_k \mid X_l)$ 검정하여 ($F = \frac{SSR(X_k, X_l) - SSR(X_l)/1}{MSE(X_k, X_l)}$ 유의성 검정) 유의하지 않으면 X_k 가 제외되고 X_l 의 유의성이 검정된다.

⑷변수가 2개 (X_k, X_l) 선택되면 $SSR(X_m \mid X_k, X_l)$ 이 가장 큰 설명 변수를 선택하고 $SSR(X_m \mid X_k, X_l)$, $SSR(X_k \mid X_m, X_l)$, $SSR(X_l \mid X_k, X_m)$ 모두에 대해 유의성 검정(F-검정)을 실시하여 유의하면 변수를 선택하고 유의하지 않는 변수들은 제외한다.

⑸유의한 설명변수가 존재하디 않을 때까지 (4)를 반복한다.

```
PROC REG DATA=LOAN;
     MODEL Loan=AGE INCOME DEBT LTV NETWTH/SELECTION=STEPWISE;
RUN;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	32606800504	16303400252	18.84	<.0001
Error	32	27686844456	865213889		
Corrected Total	34	60293644960			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-25646	21256	1259479264	1.46	0.2365
INCOME	6.93401	1.48022	18986240118	21.94	<.0001
LTV	90854	30003	7933634585	9.17	0.0048

Bounds on condition number: 1.0358, 4.143

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

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	INCOME		1	0.4092	0.4092	6.4676	22.86	<.0001
2	LTV		2	0.1316	0.5408	0.1225	9.17	0.0048

단계삽입 방법에서 default 옵션으로 SLS=0.15, SLE=0.15을 사용한다. 이를 바꾸고 싶으면 다음과 같이 하면 된다.

```
PROC REG DATA=LOAN;
     MODEL Loan=AGE INCOME DEBT LTV NETWTH
           /SELECTION=STEPWISE SLE=0.15 SLS=0.1;
RUN;
```

변수 선택은 쉽게 할 수 있도록 0.15를 사용하고 SLS는 일반적으로 사용되는 유의수준을 정하면 된다. 물론 0.05를 주로 사용하지만 0.1까지는 (신뢰수준 90%) 사회 과학에서 사용해도 무방하다.

7.3.4 방법 선택

다른 설명 변수가 주어진 경우 각 설명 변수의 설명력을 볼 수 있다는 점에서 어떤 학자들은 변수의 수가 많지 않으면 후진제거 방법을 많은 경우는 전진삽입 방법이 적합한 방법이라 한다. 통계 소프트웨어가 발달하기 전에는 계산이 복잡한 단계삽입(Stepwise) 방법이 선호되지 않았으나 이제는 이 방법이 가장 선호된다. SLE=0.2 SLS=0.1로 하자.

설명변수 수가 10개를 넘지 않는 경우는(대부분의 경우이다) 8.1절의 수작업(t-검정)이 가장 적절하다. 이유는 분석자가 자신의 지식을 이용하여 해석이 용이하거나 더 좋다고 생각하는 변수를 선택 할 수 있기 때문이다.

7.4 최적 모형 선택

서로 설명변수가 고려된 모형 중 최적 모형을 선택하고자 할 때는 $PRESS_p$ 나 수정결정계수 값을 비교하거나 새로운 자료 수집하고 모형의 예측력을 다음 방법에 의해 계산하여 각 모형을 비교하기도 한다.

$$MSPR = \frac{\sum_{i=1}^{n^*}(y_i^0 - \hat{y}_i^0)}{n^*}$$

혹은 새로운 자료로 회귀 모형을 추정하였을 때 이전 자료에서 추정된 회귀 모형과 유사하면 추정된 회귀 모형은 좋다고 판단할 수 있다. 새로운 자료 수집이 불가능한 경우에는 데이터를 splitting하여 모형 추정 데이터와 예측 데이터로 나누어 분석한다.

수정결정계수나 PRESS, AIC, SBC에 의한 최적 회귀모형 선택은 설명변수가 서로 다른 그룹을 비교할 때 사용되는 통계량이다.

HOMEWORK #10-1 (문제 조정) DUE 5월 25일(수)

FITNESS 데이터에서 종속변수는 Oxygen(산소량)이다. 나머지 변수(6개) 설명변수로 하여 다중회귀모형을 실시한다고 하자. 🖫FITNESS_IQ.xls **[SPSS 이용]**

①수작업에 의해 유의한 설명변수를 선택하시오.

②변수선택 방법 3가지(후진제거, 전진삽입, 단계삽입) 선택 방법 중 하나를 이용하여 적절한 변수를 선택하시오. (유의수준=0.1)

③유의한 변수 선택 방법 ①,②,③의 결과가 같은가? 다르다면 그 이유는 무엇인가?

④변수선택(후진제거, 전진삽입, 단계삽입) 방법, 다중공선성 문제 진단 중 어느 것을 먼저 하느냐에 따라 선택된 변수가 다른가?

⑤이상치 혹은 영향치를 진단하고 있으면 이를 제외하고 잔차 분석(8장 참고)까지 실시한 후 최종회귀모형을 추정하고 해석하시오.

HOMEWORK #10-2 (문제 조정) DUE 5월 25일(수)

종속변수 H(100,000만 가구당 살인 사건 수)에 영향을 미치는 설명변수로 G(공무원 수), M(제조업 종사자 수), W(백인 수) (연도별 데이터)만 고려한다. 🖫HOMICIDE.txt **[SAS 이용]**

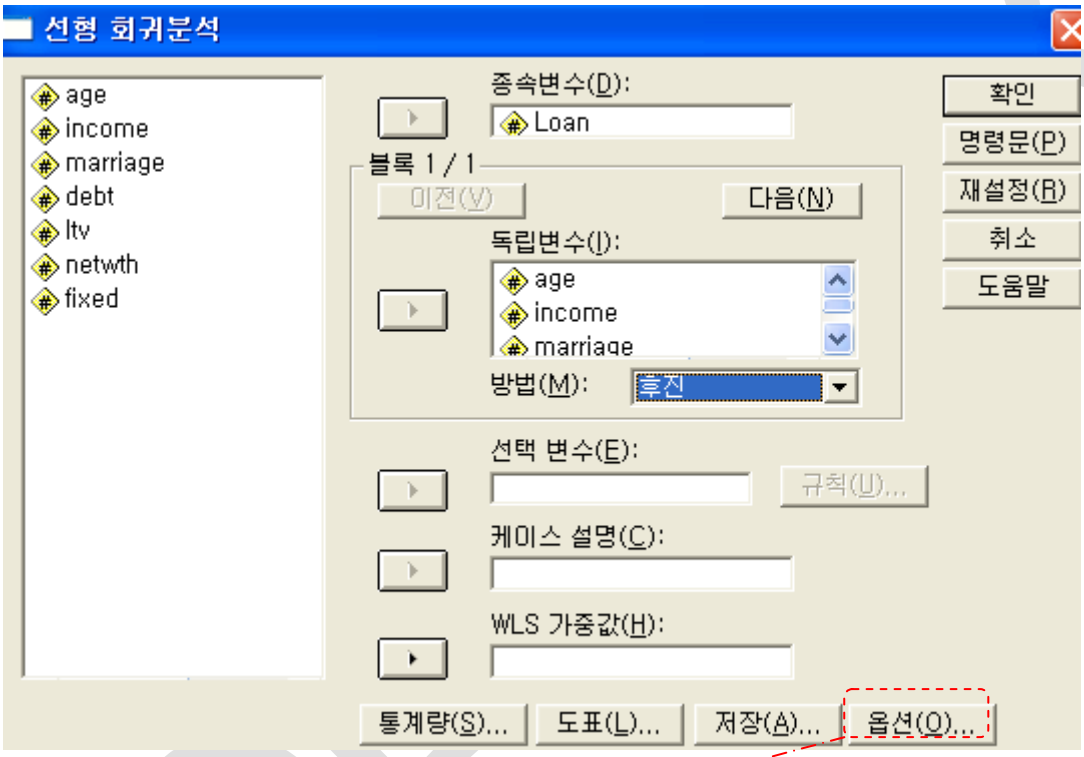
①3개 설명변수에 의해 조합할 수 있는 회귀모형을 추정하시오. 설명변수의 유의성에 상관 없이… 즉 총 7개가 생긴다. 각 회귀모형을 추정하고 모형 적합에 관련된 통계량(수정 결정 계수, Mallow C_p, PRESS, SBC, AIC)들을 계산하고 표 작성하시오.

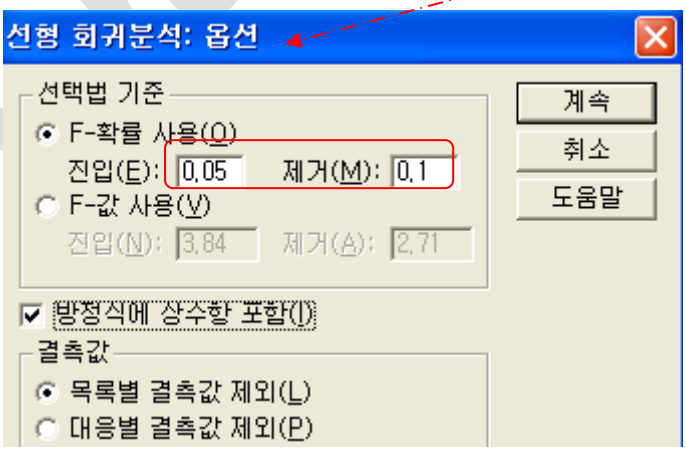
②가장 좋은 모형은? 이유는?

③영향치(이상치) 진단을 하고 잔차분석 후(8장 참고) 최종 회귀모형을 적고 해석하시오.

SPSS 에서 변수선택 하기

회귀모형 메뉴를 선택하고 종속변수와 설명변수를 설정한 후 변수 선택 방법(후진 backward, 입력 forward, 단계선택 stepwise)을 선택한다. 옵션 메뉴를 눌러 유의수준을 선택한다. SPSS는 "진입"이 "제거" 유의수준보다 반드시 크게 되어 있다. 진입을 0.1, 제거를 0.15로 사용하는 것이 일반적이다.





후진 제거 방법의 과정이 표로 출력되어 보기 편하다.

계수[a]

모형		비표준화 계수		표준화 계수	t	유의확률
		B	표준오차	베타		
1	(상수)	-41289,903	44731,961		-,923	,364
	age	168,102	677,446	,036	,248	,806
	income	5,019	2,808	,413	1,787	,085
	marriage	16581,042	15238,944	,175	1,088	,286
	debt	,012	,014	,695	,843	,407
	ltv	85159,375	35299,394	,346	2,412	,023
	netwth	-,019	,027	-,585	-,718	,479
	fixed	16669,487	13157,491	,186	1,267	,216
2	(상수)	-31966,242	23861,890		-1,340	,191
	income	5,173	2,692	,426	1,921	,065
	marriage	15341,834	14154,093	,162	1,084	,288
	debt	,011	,013	,646	,821	,419
	ltv	82289,686	32787,561	,334	2,510	,018
	netwth	-,018	,026	-,550	-,698	,491
	fixed	16414,099	12895,490	,184	1,273	,214
3	(상수)	-35754,954	23028,995		-1,553	,131
	income	4,915	2,643	,405	1,859	,073
	marriage	13326,393	13732,893	,140	,970	,340
	debt	,002	,003	,116	,570	,573
	ltv	89356,873	30905,902	,363	2,891	,007
	fixed	14528,643	12496,954	,163	1,163	,254
4	(상수)	-34409,976	22648,589		-1,519	,139
	income	6,088	1,640	,501	3,713	,001
	marriage	9224,067	11563,723	,097	,798	,431
	ltv	87043,270	30291,542	,354	2,874	,007
	fixed	12624,341	11905,878	,141	1,060	,297
5	(상수)	-28987,830	21477,356		-1,350	,187
	income	6,261	1,616	,516	3,875	,001
	ltv	88285,399	30073,499	,359	2,936	,006
	fixed	12230,770	11825,659	,137	1,034	,309
6	(상수)	-25645,718	21255,978		-1,207	,236
	income	6,934	1,480	,571	4,684	,000
	ltv	90854,114	30003,384	,369	3,028	,005

a. 종속변수: Loan

제외된 변수[f]

모형		진입-베타	t	유의확률	편상관	공선성 통계량
						공차한계
2	age	,036[a]	,248	,806	,048	,761
3	age	,018[b]	,126	,900	,024	,785
	netwth	-,550[b]	-,698	,491	-,131	,024
4	age	-,005[c]	-,035	,973	-,006	,851
	netwth	,075[c]	,365	,718	,068	,355
	debt	,116[c]	,570	,573	,105	,356
5	age	-,023[d]	-,176	,862	-,032	,879
	netwth	-,009[d]	-,046	,964	-,008	,455