Chapter 11. 회귀분석

단순회귀분석

❖ 개요

- 개념
  - 상관관계분석: 두 변수가 서로 선형관계를 가지고 있는지를 판별하는 방법
  - 회귀분석: 중속변수와 독립변수들 간의 관계를 분석하는 것
    - 단순회귀분석: 독립변수가 한 개인 경우
    - 다중회귀분석: 독립변수가 두 개 이상인 경우
- 자료
  - 종속변수: 간격척도 혹은 비율척도(계량적 자료)
  - 독립변수: 간격척도, 비율척도, 명목척도
    - 독립변수가 명목척도인 경우 독립변수를 더미변수화 하여 처리한다.
- 가정
  - 독립변수와 종속변수간의 선형적 관계
  - 오차항의 일정한 분산과 정규성
  - 오차항의 독립성

단순회귀분석

❖ 최소제곱추정법

- 최소제곱법: 편차의 제곱합을 최소화하는 모수의 값을 찾아 모수의 추정값으로 사용하는 방법
  - 편차: 관측값과 예측된 값의 차이

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} e_i^{\,2} = \sum_{i=1}^{n} (yi - \beta_0 - \beta_1 x)^2$$

- 최소제곱추정량: 편차의 제곱합을 최소화하는 모수의 추정값

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = 0 \quad \Rightarrow \quad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = 0 \quad \Rightarrow \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 x$$

  - 추정된 회귀직선

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

단순회귀분석

❖ 단순선형회귀모형에서의 추론

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 x \sim N\left(\beta_0, \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2\right)$$

- 표준화된 통계량

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma / \sqrt{S_{xx}}} \sim N(0,1) \quad \Rightarrow \quad \frac{\hat{\beta}_1 - \beta_1}{s / \sqrt{S_{xx}}} \sim t(n-2)$$

$$\frac{\hat{\beta}_0 - \beta_0}{\sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \sim N(0,1) \quad \Rightarrow \quad \frac{\hat{\beta}_0 - \beta_0}{s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \sim t(n-2)$$

단순회귀분석

- 신뢰구간

$$\beta_1 : \hat{\beta}_1 \pm t_{\alpha/2}(n-2)\frac{s}{S_{xx}}$$

$$\beta_0 : \hat{\beta}_0 \pm t_{\alpha/2}(n-2)s\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}$$

- 검정통계량
  - $H_0 : \beta_1 = \beta_{10}$ 일 때, 검정통계량: $t = \dfrac{\hat{\beta}_1 - \beta_1}{s/\sqrt{S_{xx}}} \sim t(n-2)$
  - $H_0 : \beta_0 = \beta_{00}$ 일 때, 검정통계량: $t = \dfrac{\hat{\beta}_0 - \beta_0}{s\sqrt{\dfrac{1}{n}+\dfrac{\bar{x}^2}{S_{xx}}}} \sim t(n-2)$

단순회귀분석

❖ 예제: 광고비(독립변수)와 매출액(종속변수)

단순 회귀분석

```
DATA adsales;
INPUT company adver sales @@;
CARDS;
01 11 23 02 19 32 03 23 36 04 26 46 05 56 93
06 62 99 07 29 49 08 30 50 09 38 65 10 39 70
11 46 71 12 49 89
;
RUN;

PROC REG DATA=adsales;
MODEL sales=adver;
PLOT sales*adver;
RUN;
```

단순회귀분석

- 적합된 회귀식
  - SALES = 3.28480 + 1.59716 * ADVER

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6695.27457	6695.27457	455.54	<.0001
Error	10	146.97543	14.69754		
Corrected Total	11	6842.25000			

Root MSE	3.83374	R-Square	0.9785
Dependent Mean	60.25000	Adj R-Sq	0.9764
Coeff Var	6.36305		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.28480	2.88935	1.14	0.2821
ADVER	1	1.59716	0.07483	21.34	<.0001

단순회귀분석

❖ 종속변수에 대한 예측

$$E(y \mid x = x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 \sim N\left(\hat{\beta}_0 + \hat{\beta}_1 x_0, \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) \sigma^2 \right)$$

- σ^2 를 모르는 경우

$$\frac{\left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right) - \left(\beta_0 + \beta_1 x_0\right)}{s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t(n-2)$$

- 종속변수에 대한 신뢰구간

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right) \pm t_{\alpha/2}(n-2) s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

단순회귀분석

❖ 예제: 광고비(독립변수)와 매출액(종속변수)

단순 회귀분석

```
DATA adsales;
INPUT company adver sales @@;
CARDS;
01 11 23 02 19 32 03 23 36 04 26 46 05 56 93
06 62 99 07 29 49 08 30 50 09 38 65 10 39 70
11 46 71 12 49 89
;
RUN;

PROC REG DATA=adsales;
MODEL sales=adver/P CLM ALPHA=0.01;
RUN;
```

단순회귀분석

```
PROC REG DATA=A;
MODEL SALES=ADVER / P CLM ALPHA=0.01;
RUN;QUIT;
```

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	99% CL Mean		Residual
1	23.0000	20.8535	2.1522	14.0326	27.6744	2.1465
2	32.0000	33.6307	1.6674	28.3462	38.9153	-1.6307
3	36.0000	40.0194	1.4571	35.4013	44.6374	-4.0194
4	46.0000	44.8108	1.3221	40.6206	49.0011	1.1892
5	93.0000	92.7255	1.8815	86.7625	98.6884	0.2745
6	99.0000	102.3084	2.2601	95.1456	109.4712	-3.3084
7	49.0000	49.6023	1.2139	45.7550	53.4496	-0.6023
8	50.0000	51.1995	1.1852	47.4434	54.9556	-1.1995
9	65.0000	63.9767	1.1204	60.4259	67.5275	1.0233
10	70.0000	65.5739	1.1345	61.9784	69.1693	4.4261
11	71.0000	76.7539	1.3501	72.4752	81.0327	-5.7539
12	89.0000	81.5454	1.4901	76.8230	86.2678	7.4546

Sum of Residuals	0
Sum of Squared Residuals	146.97543
Predicted Residual SS (PRESS)	210.84863

단순회귀분석

❖ 잔차분석

- 오차에 대한 가정
  - $E(\varepsilon_i) = 0$
  - 독립성: ε_i 는 독립
  - 등분산성: $Var(\varepsilon_i) = \sigma^2$
  - 정규성: $\varepsilon_i \sim N(0,1)$
- 잔차

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

- 표준화 잔차
  - 표본의 크기가 충분히 클 때, 근사적으로 서로 독립이며 정규분포를 따른다.
  - 0을 중심으로 -2와 2사이에 랜덤하게 분포

단순회귀분석

❖ 예제: 광고비(독립변수)와 매출액(종속변수)

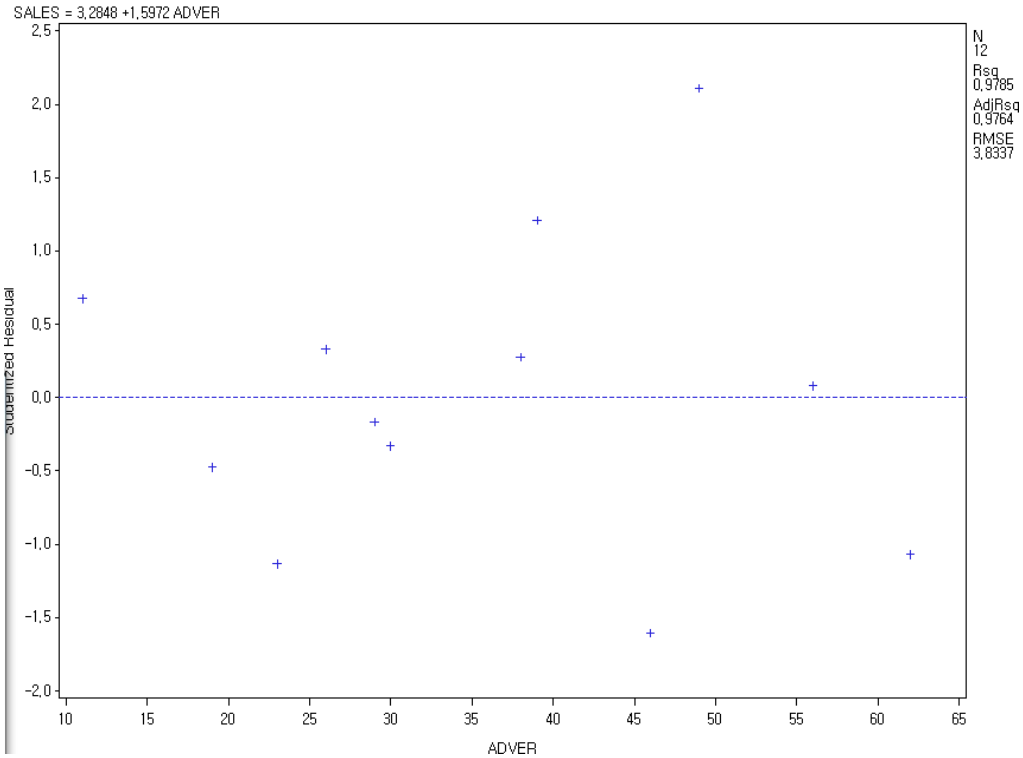
표준화 잔차와 잔차도표 출력

```
DATA adsales;
INPUT company adver sales @@;
CARDS;
01 11 23 02 19 32 03 23 36 04 26 46 05 56 93
06 62 99 07 29 49 08 30 50 09 38 65 10 39 70
11 46 71 12 49 89
;
RUN;

PROC REG DATA=adsales GRAPHICS;
MODEL sales=adver/R;
OUTPUT OUT=regout STUDENT=std_r;
PLOT STUDENT.*adver;
RUN;
```

단순회귀분석

Student Residual	-2-1 0 1 2	Cook's D
0.677	*	0.105
-0.472		0.026
-1.133	**	0.108
0.330		0.007
0.0822		0.001
-1.068	**	0.304
-0.166		0.002
-0.329		0.006
0.279		0.004
1.209	**	0.070
-1.604	***	0.182
2.110	****	0.396



단순회귀분석

- 잔차의 정규성 검토

히스토그램과 정규확률도표의 출력

```
DATA adsales;
INPUT company adver sales @@;
CARDS;
01 11 23 02 19 32 03 23 36 04 26 46 05 56 93
06 62 99 07 29 49 08 30 50 09 38 65 10 39 70
11 46 71 12 49 89
;
RUN;

PROC UNIVARIATE DATA=regout;
VAR std_r;
HISTOGRAM std_r/NORMAL;
RUN;
```

정규 분포에 대한 모수

모수	심볼	추정값
평균	Mu	-0.00709
표준편차	Sigma	1.041073

정규 분포에 대한 적합도 검정

검정		------통계량------		-------p-값--------
Kolmogorov-Smirnov	D	0.12288172	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.02181102	Pr > W-Sq	>0.250
Anderson-Darling	A-Sq	0.16412536	Pr > A-Sq	>0.250

단순회귀분석

- 독립성 검토: Durbin-Watson 검정통계량 이용
  - DW 값이 2에 가까울수록 독립

Durbin-Watson 검정

```
DATA adsales;
INPUT company adver sales @@;
CARDS;
01 11 23 02 19 32 03 23 36 04 26 46 05 56 93
06 62 99 07 29 49 08 30 50 09 38 65 10 39 70
11 46 71 12 49 89
;
RUN;

PROC REG DATA=adsales;
MODEL sales=adver/DW;
RUN;
```

```
The REG Procedure
Model: MODEL1
Dependent Variable: SALES

Durbin-Watson D                  2.470
Number of Observations              12
1st Order Autocorrelation       -0.440
```

단순회귀분석

- ❖ 선형관계의 강도: 결정계수
  - 선형모형이 어느정도 적합한가의 측도
  - Y의 총 변동 중 선형회귀모형에 의하여 설명되는 변동부분

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- ❖ 분산분석
  - 선형 모형에 사용되는 계수들이 모두 0인지를 검정
  - 귀무가설: 모든 계수 = 0 vs 대립가설: 0이 아닌 계수가 존재

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6695.27457	6695.27457	455.54	<.0001
Error	10	146.97543	14.69754		
Corrected Total	11	6842.25000			

Root MSE	3.83374	R-Square	0.9785
Dependent Mean	60.25000	Adj R-Sq	0.9764
Coeff Var	6.36305		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.28480	2.88935	1.14	0.2821
ADVER	1	1.59716	0.07483	21.34	<.0001

연습문제1

- 음주운전자의 혈액 중 알코올 농도를 측정하기 위해 교통경찰관이 사용하는 디지털 측정기의 신뢰성을 알아보려고 한다. 15명의 음주 운전자에 대해 혈액채취(y)와 디지털 측정기(x)에 의해 각각 알코올 농도를 측정하여 다음과 같은 데이터를 얻었다. X에 의한 y의 회귀직선식을 구하고 잔차분석을 실시하여라.

X	Y	X	Y	X	Y
0.150	0.154	0.090	0.082	0.110	0.078
0.100	0.085	0.090	0.072	0.120	0.097
0.900	0.079	0.090	0.080	0.100	0.088
0.140	0.144	0.095	0.090	0.060	0.053
0.080	0.078	0.040	0.050	0.080	0.072

다중회귀분석

❖ 개요

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \cdots + \beta_k x_k + \varepsilon$$

- 다중회귀분석의 개념과 추정방식
  - 개념
    - 두 개 이상의 독립변수들과 하나의 종속변수의 관계를 분석하는 기법
  - 추정방식
    - 동시입력방식
      - 연구자가 고려하는 모든 독립변수들을 한꺼번에 포함하여 분석하는 방법
      - 독립변수들이 동시에 종속변수를 설명하는 정도를 알 수 있음
    - 단계선택방식
      - 종속변수에 영향력이 있는 변수들만을 회귀식에 포함시키는 방법
      - 영향력이 높은 변수의 순으로 회귀식에 포함
      - 포함된 독립변수들도 나중에 들어오는 변수 때문에 설명력이 낮아지면 회귀식에서 제거되어 짐
      - 설명력이 어느 정도 이상 되는 변수들로만 구성된 회귀식을 발견하는데 유용
- 자료 와 가정
  - 단순회귀분석과 동일

다중회귀분석

- 예제: 에어로빅 적합성을 알아보기 위해 31명으로부터 oxygen(산소섭취율, 종속변수), age(나이), weight(체중), runtime(1.5마일을 주행하는데 소요되는 시간), rstpulse(휴식중 맥박수), runpulse(주행중의 맥박수), maxpulse(주행중의 최대맥박수)을 측정한 자료

다중 선형 회귀분석

```
DATA fitness;
INFILE ‘C:\fitness.txt’;
INPUT oxygen age weight runtime rstpulse runpulse maxpulse @@;
LABEL oxygen='산소섭취율' age='나이' weight='체중' runtime='1.5마일주행시간’ rstpulse='휴식중맥박수' runpulse='주행중맥박수' maxpulse='주행중최대맥박수';
RUN;

PROC REG DATA=adsales;
MODEL oxygen=age weight runtime rstpulse runpulse maxpulse;
RUN;
```

다중회귀분석

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	722.54361	120.42393	22.43	<.0001
Error	24	128.83794	5.36825		
Corrected Total	30	851.38154			

Root MSE	2.31695	R-Square	0.8487
Dependent Mean	47.37581	Adj R-Sq	0.8108
Coeff Var	4.89057		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	102.93448	12.40326	8.30	<.0001
AGE	나이	1	-0.22697	0.09984	-2.27	0.0322
WEIGHT	체중	1	-0.07418	0.05459	-1.36	0.1869
RUNTIME	1.5마일주행시간	1	-2.62865	0.38456	-6.84	<.0001
RSTPULSE	휴식중맥박수	1	-0.02153	0.06605	-0.33	0.7473
RUNPULSE	주행중맥박수	1	-0.36963	0.11985	-3.08	0.0051
MAXPULSE	주행중최대맥박수	1	0.30322	0.13650	2.22	0.0360

다중회귀분석

- 변수선택
  - SELECTION=
    - FOREWARD:하나씩 들어오기
      - SLENTRY=0.15
    - BACKWARD:하나씩 나가기
      - SLSTAY=0.15
    - STEPWISE: FOREWARD+BACKWARD
      - SLENTRY=,SLSTAY=

다중 선형 회귀분석

```
DATA fitness;
INFILE ‘C:\fitness.txt’;
INPUT oxygen age weight runtime rstpulse runpulse maxpulse @@;
LABEL oxygen='산소섭취율' age='나이' weight='체중' runtime='1.5마일주행시간’ rstpulse='휴식중맥박수' runpulse='주행중맥박수' maxpulse='주행중최대맥박수';
RUN;

PROC REG DATA=adsales;
MODEL oxygen=age weight runtime rstpulse runpulse maxpulse/SELECTION=STEPWISE;
RUN;
```

다중회귀분석

Stepwise Selection: Step 4

Variable MAXPULSE Entered: R-Square = 0.8368 and C(p) = 4.8800

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	712.45153	178.11288	33.33	<.0001
Error	26	138.93002	5.34346		
Corrected Total	30	851.38154			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	98.14789	11.78569	370.57373	69.35	<.0001
AGE	-0.19773	0.09564	22.84231	4.27	0.0488
RUNTIME	-2.76758	0.34054	352.93570	66.05	<.0001
RUNPULSE	-0.34811	0.11750	46.90089	8.78	0.0064
MAXPULSE	0.27051	0.13362	21.90067	4.10	0.0533

Bounds on condition number: 8.4182, 76.851

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.

다중회귀분석

❖ 다중공선성

- 독립변수들 간의 완전한 또는 거의 완전한 선형종속의 관계를 의미
- 탐색방법
  - 분산확대인자(variance inflation: VIF):10보다 큰 경우 다중공선성이 있음
    - 다중공선성이 있는 경우: 변수선택, 능형회귀,주성분회귀 등..
  - 상태지수(condition index):100보다 큰 경우

다중 공선성진단

```
DATA fitness;
INFILE 'C:\fitness.txt';
INPUT oxygen age weight runtime rstpulse runpulse maxpulse @@;
LABEL oxygen='산소섭취율' age='나이' weight='체중' runtime='1.5마일주행시간' rstpulse='휴식중맥박수' runpulse='주행중맥박수' maxpulse='주행중최대맥박수';
RUN;

PROC REG DATA=adsales;
MODEL oxygen=age weight runtime rstpulse runpulse maxpulse/VIF COLLIN;
RUN;
```

다중회귀분석

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	102.93448	12.40326	8.30	<.0001	0
AGE	나이	1	-0.22697	0.09984	-2.27	0.0322	1.51284
WEIGHT	체중	1	-0.07418	0.05459	-1.36	0.1869	1.15533
RUNTIME	1.5마일주행시간	1	-2.62865	0.38456	-6.84	<.0001	1.59087
RSTPULSE	휴식중맥박수	1	-0.02153	0.06605	-0.33	0.7473	1.41559
RUNPULSE	주행중맥박수	1	-0.36963	0.11985	-3.08	0.0051	8.43727
MAXPULSE	주행중최대맥박수	1	0.30322	0.13650	2.22	0.0360	8.74385

Collinearity Diagnostics

Number	Eigenvalue	Condition Index
1	6.94991	1.00000
2	0.01868	19.29087
3	0.01503	21.50072
4	0.00911	27.62115
5	0.00607	33.82918
6	0.00102	82.63757
7	0.00017947	196.78560

다중회귀분석

❖ 최종모형

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	690.55086	230.18362	38.64	<.0001
Error	27	160.83069	5.95669		
Corrected Total	30	851.38154			

Root MSE	2.44063	R-Square	0.8111
Dependent Mean	47.37581	Adj R-Sq	0.7901
Coeff Var	5.15165		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	111.71806	10.23509	10.92	<.0001	0
AGE	나이	1	-0.25640	0.09623	-2.66	0.0129	1.26661
RUNTIME	1.5마일주행시간	1	-2.82538	0.35828	-7.89	<.0001	1.24444
RUNPULSE	주행중맥박수	1	-0.13091	0.05059	-2.59	0.0154	1.35476

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation			
			Intercept	AGE	RUNTIME	RUNPULSE
1	3.97790	1.00000	0.00011565	0.00056585	0.00082368	0.00016363
2	0.01183	18.33958	0.00296	0.38305	0.49678	0.00697
3	0.00919	20.80033	0.03198	0.19423	0.42448	0.09749
4	0.00108	60.60078	0.96495	0.42215	0.07792	0.89538

연습문제 2

- 환경변화가 혈압에 미치는 장기적인 변화를 연구하기 위하여 안데스산맥의 고지대에서 도시로 이주해온 인디오들로부터 여러 특성들을 측정하였다. 다음의 데이터는 그 중 최고혈압(y)과 신체적 특성들인 나이(x1), 이주 후 경과기간(x2,단위:년),몸무게(x3,단위:kg),복부피부두께(x4,단위:mm)등에 관한 결과이다. 다중 회귀분석을 수행하여라.

21	1	71	12.7	170 22	6	56.5	8	120
24	5	56	4.3	125 24	1	61	4.3	148
25	1	65	20.7	140 27	19	62	5.7	106
28	5	53	8	120 28	25	53	0	108
31	6	65	10	124 32	13	57	6	134
33	13	66.5	8.3	116 33	10	59.1	10.3	114
34	15	64	7	130 35	18	69.5	7	118
35	2	64	6.7	138 36	12	56.5	11.7	134
36	15	57	6	120 37	16	55	7	120
37	17	57	11.7	114 38	10	58	13	124
38	18	59.5	7.7	114 38	11	61	4	136
38	11	57	3	126 39	21	57.5	5	124
39	24	74	15.7	128 39	14	72	13.3	134
41	25	62.5	8	112 41	32	68	11.3	128
41	5	63.4	13.7	134 42	12	68	10.7	128
43	25	69	6	140 43	26	73	5.7	138
43	10	64	7	118 44	19	65	7.7	110
44	18	71	4.3	142 45	10	60.2	3.3	134
47	1	55	4	116 50	43	70	11.7	132
54	40	87	11.3	152				