Matrix

▪정의

- A collection of numbers arranged into a fixed number of rows and columns

▪형태

- 차수 (nxp)인 행렬 matrix
  - 원소 $\{x_{ij}\}$
- 열 벡터 column vector
- 행 벡터 row vector
- 스칼라 scalar
- 변수 벡터 Variable vector
  - 데이터 벡터 data vector

←측정변수(p)→

↑ 개체 (n) ↓

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ & \cdots\cdots\cdots\cdots & & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

$$\underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

▪Special matrix

- 정방행렬 square matrix
- 대각행렬 Diagonal matrix
- 항등행렬 identity matrix
  - 항등벡터
  - 1 벡터, 1 행렬
- 영 행렬 null matrix

▪예제 데이터

- 학생 3명의 가족 수와 외식 회수 조사

$$X = \begin{bmatrix} 3 & 0 \\ 2 & 2 \\ 4 & 1 \end{bmatrix}$$

▪예제 행렬

$$A = \begin{bmatrix} 2 & -1 & 2 \\ 3 & 0 & 1 \end{bmatrix},\ B = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix},\ C = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}, \underline{D} = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$

▪행렬 연산 Operation

- 동일
  - 더하기 addition: 대응 원소의 합
  - 차수 동일
- 연산 적합 Conformable
- 대각합 trace
- Augmented 증가함수
- 전치 transpose
- 곱하기 multiplication
  - 앞의 열 벡터와 뒤 행 벡터를 곱한다. 앞 열차수=뒤 행 차수
  - 결과: (앞 행 차수)X(뒤 열 차수)

$tr(B) = 2, tr(C) = 5$

$X \mid D$

$$X' = \begin{bmatrix} 3 & 2 & 4 \\ 0 & 2 & 1 \end{bmatrix}$$

$$A\underline{D} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}, \underline{D}'X = \begin{bmatrix} -1 & -1 \end{bmatrix}$$

http://wolfpack.hnu.ac.kr

Operation (cont.)

▪ 합, 전치, 곱 연산 성질
- $(A+B)'=A'+B'$
- $(AB)'=B'A'$
- $tr(A+B)=tr(A)+tr(B)$
- $tr(AB)=tr(BA)$
- $(A+B)+C=A+(B+C)$ 결합법칙 association
- $A(B+C)=AB+AC$ 배분법칙 distribution

▪ 대칭행렬 symmetric matrix
- 만약 $A'=A$, A은 대칭행렬
- X'X는 대칭행렬
  - X는 데이터 행렬이다.

▪ 멱등행렬 idempotent matrix
- 만약 MM=M이면 M은 멱등행렬
- If M is idempotent, $M^K=M$.

▪ 모든 원소가 1인 행렬
- (모든 원소가 1인 행벡터)*(모든 원소가 1인 열벡터)
- (모든 원소가 1인 열벡터)*(모든 원소가 1인 행벡터)

▪ 역행렬 inverse matrix
- 만약 $AA^{-1}=I$, $A^{-1}A=I$, A^{-1}를 역행렬이라 한다
- 행렬식 determinant |A| $C=\begin{bmatrix}1 & 3\\ 2 & 4\end{bmatrix} \Rightarrow |C| = 4-6=-2$
  - Minor(소 행렬식) $|M_{ij}|$, cofactor (여인자)

$$A=\begin{bmatrix}1 & 2 & 3\\ 4 & 5 & 7\\ 8 & 9 & 10\end{bmatrix} \qquad |A| = \sum_{i=1}^{n} a_{ij}(-1)^{i+j}|M_{ij}| = \sum_{j=1}^{n} a_{ij}(-1)^{i+j}|M_{ij}|$$

$$|A| = 1(-1)^{1+1}\begin{vmatrix}5 & 7\\ 9 & 10\end{vmatrix} + 2(-1)^{1+2}\begin{vmatrix}4 & 7\\ 8 & 10\end{vmatrix} + 3(-1)^{1+3}\begin{vmatrix}4 & 5\\ 8 & 9\end{vmatrix} = 7$$

▪ 행렬식 성질
- $|A'|=|A|$, $|AB|=|A||B|=|BA|$
- $|A^2|=|A|^2$
- 두 행(열)이 동일하면 행렬식은 0이다.
- 한 행이 다른 행의 선형 함수로 표현되면 0이다.

http://wolfpack.hnu.ac.kr

Inverse Matrix (cont.)

▪역행렬 계산 정의

$$A^{-1} = \frac{1}{|A|} adjoint(A)$$

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \rightarrow |A| = 1\times4 - 2\times3 = -2$$

$$adj(A) = \begin{bmatrix} (-1)^{1+1}4 & (-1)^{1+2}3 \\ (-1)^{2+1}2 & (-1)^{2+2}1 \end{bmatrix}' = \begin{bmatrix} 4 & -3 \\ -2 & 1 \end{bmatrix}' = \begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$$

▪역행렬 성질

$$\mathrm{A} \equiv \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix},$$

- 역행렬은 unique하다.
- $|A^{-1}|=1/|A|$
- $(A')^{-1}=(A^{-1})'$
- $(AB)^{-1}=B^{-1}A^{-1}$

▪계수(rank) 정의

- 아래 식이 모든 a_i가 0일 때만 만족한다면 x_i들은 Linearly Independent

$$a_1\underline{x}_1 + a_2\underline{x}_2 + \ldots + a_p\underline{x}_p = \underline{0}$$

- 계수: 행렬 A의 LIN 행(혹은 열)의 수
- Full rank: 행렬 A의 차수와 rank가 같을 때

▪행렬 A에 대하여

▪ 역 행렬이 존재한다. ▪ full-rank이다. rank(A)=n ▪ A는 non-singular이다. ▪ $\|A\|\neq0$ ▪ Ax=b 의 해가 존재	▪ 역 행렬이 존재하지 않는다 ▪ full-rank아니다. rank(A)<n ▪ A는 singular이다. ▪ $\|A\|=0$ ▪ Ax=b 의 해가 존재하지 않음.

▪고유치 eigen value & 고유벡터 eigen vector

- $|A-\lambda I|=0$을 만족하는 λ를 고유치라 한다.
  - 고유치는 실수
  - 공분산, 상관계수 행렬과 같이 positive definite matrix의 고유치 의 개수는 행렬의 차수와 동일하자.
- $Ae=\lambda_i e$를 만족하는 벡터 e를 고유벡터라 한다. 무수히 많이 존재한다.

$$\mathrm{A}^{-1} = \frac{1}{|\mathrm{A}|}\begin{bmatrix} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} & \begin{vmatrix} a_{13} & a_{12} \\ a_{33} & a_{32} \end{vmatrix} & \begin{vmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{vmatrix} \\ \begin{vmatrix} a_{23} & a_{21} \\ a_{33} & a_{31} \end{vmatrix} & \begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix} & \begin{vmatrix} a_{13} & a_{11} \\ a_{23} & a_{21} \end{vmatrix} \\ \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix} & \begin{vmatrix} a_{12} & a_{11} \\ a_{32} & a_{31} \end{vmatrix} & \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \end{bmatrix}.$$

Matrix OPERATION in SAS and R

▪연립방정식 해

$$2u-2v-w=5$$
$$u+v-2w=1$$
$$u-w=4$$

$$A\underline{x}=\underline{b} \Leftrightarrow \begin{bmatrix} 2 & 2 & -1 \\ 1 & 1 & -2 \\ 1 & 0 & -1 \end{bmatrix}\begin{bmatrix} u \\ v \\ w \end{bmatrix}=\begin{bmatrix} 5 \\ 1 \\ 4 \end{bmatrix}$$

▪Interactive Matrix Language

```
PROC IML;
    A={2 2 -1, 1 1 -2, 1 0 -1};
    A1=A`;
    B=DET(A);
    B1=DET(A1);
    C=INV(A);
    CALL EIGEN(M,E,A);
PRINT A  A1, B B1, C, M E;
RUN;
```

```
        A                              A1

        2        2        -1        2        1        1
        1        1        -2        2        1        0
        1        0        -1       -1       -2       -1
                           B        B1

                          -3        -3
                            C

            0.3333333 -0.666667          1
            0.3333333 0.3333333         -1
            0.3333333 -0.666667          0
              M                E

  2.3027756            0 -0.919272 0.8728716 0.2093505
          1            0 -0.278333 -0.218218 -0.691438
  -1.302776            0 -0.278333 0.4364358 -0.691438
```

▪In R

```
> a<-matrix(c(2,2,-1,1,1,-2,1,0,-1),3,3)
> a
     [,1] [,2] [,3]
[1,]    2    1    1
[2,]    2    1    0
[3,]   -1   -2   -1
> t(a)
     [,1] [,2] [,3]
[1,]    2    2   -1
[2,]    1    1   -2
[3,]    1    0   -1
> det(a)
[1] -3
> solve(a)
           [,1]       [,2]       [,3]
[1,]  0.3333333  0.3333333  0.3333333
[2,] -0.6666667  0.3333333 -0.6666667
[3,]  1.0000000 -1.0000000  0.0000000
> a%*%solve(a)
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    1
> eigen(a)
$values
[1]  2.302776 -1.302776  1.000000

$vectors
           [,1]       [,2]          [,3]
[1,] -0.4528821 -0.3608454 -2.498669e-16
[2,] -0.6952572  0.3134004  7.071068e-01
[3,]  0.5581355  0.8783910 -7.071068e-01
```

Matrix Application

▪Data matrix

▪ 평균 벡터 mean vector

$$\underline{x}(\underline{\mu}) = (\frac{1}{n})1_n' X_{n\times p} \qquad X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ & \cdots\cdots & \cdots\cdots & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

$$\overline{X}_i = \frac{\sum_j X_{ji}}{n}$$

▪공분산 행렬 covariance matrix

$$\sigma_{jk} = \mathrm{cov}(x_j \ \ x_k) = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_{ij} - \overline{x}_i)(x_{ik} - \overline{x}_k)$$

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \ldots \sigma_{1p} \\ \sigma_{21} & \sigma_{22} \ldots \sigma_{2p} \\ & \vdots \\ \sigma_{p1} & \sigma_{p2} \ldots \sigma_{pp} \end{bmatrix} \qquad Corr(x_i, x_j) = \frac{COV(x_i, x_j)}{\sqrt{V(x_i)}\sqrt{V(x_j)}}$$

$$S(\Sigma) = \frac{1}{n-1}(X - 1\underline{x}')'(X - 1\underline{x}')$$

▪변수 벡터 Variable vector

▪ 평균

$$\underline{\mu} = E(\underline{x}) = \begin{bmatrix} E(x_1) \\ E(x_2) \\ \vdots \\ E(x_p) \end{bmatrix} \qquad \underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

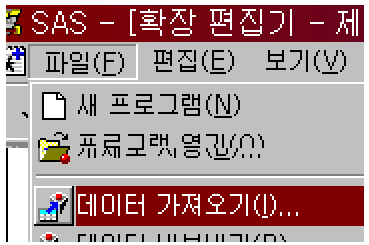
▪ 공분산

$$\Sigma = E(\underline{x} - E(\underline{x}))(\underline{x} - E(\underline{x}))'$$

▪Data 예제 LPGAtour.XLS

▪ Read Data in SAS

- WORK 라이브러리에
- LPGA 이름으로 저장



▪ Read Data in SAS

- 우선 CSV 포멧으로 저장

In SAS

- 비거리, 페어웨이 적중률, 그린 적중률
  - PROC 이용
  - IML 이용

```
PROC CORR DATA=LPGA COV;
   VAR DRIVING_DISTANCE FAIRWAYS GREENS;
RUN;

DATA LPGA1;
   SET LPGA;
   KEEP DRIVING_DISTANCE FAIRWAYS GREENS;
RUN;

PROC IML;
   USE LPGA1;
   READ ALL INTO X;
   I=j(20,1);
   M=(1/NROW(I))*I`*X;
   J=j(20,1);
   DEV=X-J*M;
   COV=(1/(NROW(I)-1))*DEV`*DEV;
   PRINT M , DEV, COV;
RUN;
```

- 고유치, 고유벡터 구하기
  - 고유치 m (차수만큼 존재)
  - 고유벡터 e

```
PROC IML;
   X={9 2, 2 4};
   CALL EIGEN(M,E,X);
   PRINT M, E;
RUN;
QUIT;
```

- 결과

공분산행렬, DF = 19

	Driving_ Distance	Fairways	Greens
Driving Distance	91.82842105	-39.71157895	17.08000000
Fairways	-39.71157895	35.23115789	1.70642105
Greens	17.08000000	1.70642105	13.11313158

단순 통계량

N	평균	표준편차	합	최소값	최대값
20	245.90000	9.58271	4918	227.20000	261.20000
20	69.32000	5.93558	1386	56.40000	79.10000
20	69.44500	3.62121	1389	63.30000	78.60000

M

245.9	69.32	69.445

DEV

3.1	6.08	3.755
8.7	2.78	5.655
10.1	-0.02	9.155
3.5	-2.42	0.955
-6.5	1.08	-0.845
5.7	-0.32	0.355
-2.4	-2.72	-3.545

COV

91.828421	-39.71158	17.08
-39.71158	35.231158	1.7064211
17.08	1.7064211	13.113132

M

9.7015621
3.2984379

E

0.9436283	-0.331007
0.3310069	0.9436283

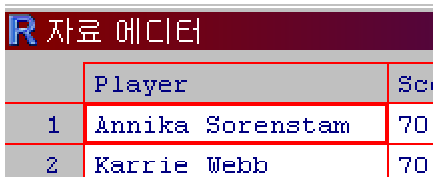
In R

■비거리, 페어웨이 적중률, 그린 적중률

```
> lpga<-read.table("LPGATour.csv",sep=",",
+ header=T,na.string=".")
> lpga
              Player Scoring.Avg. Driving.Dista
1     Annika Sorenstam        70.04            24
2          Karrie Webb        70.00            25
3        Kelly Robbins        70.35            25
4        Chris Johnson        70.84            24
```

■Data 수정

```
> fix(lpga)
```



■Data Subset

```
> lpga1<-cbind(lpga$Driving.Distance,
+ lpga$Fairways,lpga$Greens)
> lpga1
      [,1] [,2] [,3]
 [1,] 249.0 75.4 73.2
 [2,] 254.6 72.1 75.1
 [3,] 256.0 69.3 78.6
 [4,] 249.4 66.9 70.4
```

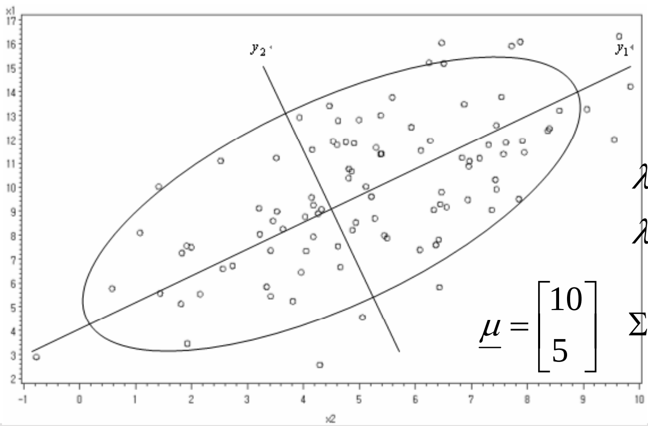
■평균, 공분산 행렬, 고유치 구하기 구하기

```
> (1/nrow(lpga1))%*%(t(matrix(1,20,1))%*%lpga1)
      [,1]  [,2]   [,3]
[1,] 245.9 69.32 69.445

> M<-(1/nrow(lpga1))%*%(t(matrix(1,20,1))%*%lpga1)

> COV<-(1/(nrow(lpga1)-1))*
+ (t(lpga1-matrix(1,20,1)%*%M)%*%
+ (lpga1-matrix(1,20,1)%*%M))
> COV
          [,1]       [,2]      [,3]
[1,]  91.82842 -39.711579 17.080000
[2,] -39.71158  35.231158  1.706421
[3,]  17.08000   1.706421 13.113132

> eigen(COV)
$values
[1] 114.358942  22.335573   3.478196

$vectors
           [,1]       [,2]       [,3]
[1,]  0.8859250 -0.2649759 -0.3806897
[2,] -0.4415535 -0.7330991 -0.5172970
[3,]  0.1420120 -0.6263813  0.7664719
```

HW #2-1 Due 2008.10.13 (Mon)

- **LPGA 데이터에서**
  - 4개 변수(DRIVING_DISTANCE FAIRWAYS GREENS SAND_SAVES)를 데이터 행렬 X라 하자.
  - In SAS/IML, R에서 다음 작업을 하시오.
    - 데이터 행렬 X의 (X'X)가 대칭행렬임을 보이시오.
    - 행렬 X의 표본 평균 벡터, 표본 공분산 행렬을 구하시오.
    - 표본 공분산 행렬을 이용하여 고유치, 고유벡터를 구하시오.
    - 종속변수(벡터 y): Scoring Average 스코어, 설명변수(행렬X): 비거리, 그린 적중률, 페어웨이 적중률 회귀모형의 OLS 추정치를 구하시오.

$$\underline{y} = X\underline{\beta} + \underline{e} \sim N(\underline{0}, \sigma^2 I) \quad OLS: \hat{\underline{\beta}} = \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = (X'X)^{-1}X'\underline{y}$$



$$\lambda_1 = 9.7 \qquad \lambda_2 = 3.2$$

$$\underline{\mu} = \begin{bmatrix} 10 \\ 5 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 9 & 2 \\ 2 & 4 \end{bmatrix}$$

```
PROC PRINCOMP DATA=LPGA COV;
   VAR DRIVING_DISTANCE FAIRWAYS GREENS SAND_SAVES;
RUN;
```

Simple Statistics

	Driving_ Distance	Fairways	Greens	Sand_Saves
Mean	245.9000000	69.32000000	69.44500000	42.21500000
StD	9.5827147	5.93558404	3.62120582	5.02365719

Covariance Matrix

	Driving_ Distance	Fairways	Greens	Sand_Saves
Driving Distance	91.82842105	-39.71157895	17.08000000	3.18789474
Fairways	-39.71157895	35.23115789	1.70642105	2.11968421
Greens	17.08000000	1.70642105	13.11313158	3.46507895
Sand Saves	3.18789474	2.11968421	3.46507895	25.23713158

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	114.422636	85.882181	0.6918	0.6918
2	28.540455	9.566098	0.1725	0.8643
3	18.974357	15.501963	0.1147	0.9790
4	3.472394		0.0210	1.0000

Eigenvectors

	Prin1	Prin2	Prin3	Prin4
Driving Distance	0.885927	0.133061	0.231532	0.379237
Fairways	-.440465	0.438673	0.590719	0.514400
Greens	0.142856	0.376642	0.496429	-.768954
Sand Saves	0.026749	0.804985	-.592454	0.016777

```
PROC IML;
   X={9 2, 2 4};
   CALL EIGEN(M,E,X);
   PRINT M, E;
RUN;
QUIT;
```

```
M
9.7015621
3.2984379
E
0.9436283 -0.331007
0.3310069 0.9436283
```

HW #2-2 Due 2008.10.13 (Mon)

- Generating Bivariate Normal Dist. (n=100)
  - Generating a BN form the following situation
  - 두 변수의 산점도를 그리시오.
  - 모집단 공분산 행렬로부터 고유치, 고유벡터를 구하시오.
  - 표본 평균 벡터와 표본 공분산 행렬을 구하시오.
  - 표본 공분산 행렬로부터 고유치, 고유벡터를 구하시오.

    - (1) $\underline{\mu} = \begin{bmatrix} 6 \\ 5 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 9 & 5 \\ 5 & 4 \end{bmatrix}$
    - (2) $\underline{\mu} = \begin{bmatrix} 6 \\ 5 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 9 & 1 \\ 1 & 4 \end{bmatrix}$
    - (3) $\underline{\mu} = \begin{bmatrix} 9 \\ 30 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 9 & 13 \\ 13 & 25 \end{bmatrix}$

  - 위의 작업을...
    - SAS/IML 이용하여 하시오.
    - R을 이용하여 하시오.

```
> plot(lpga$Fairways,lpga$Greens)
```

```
PROC GPLOT DATA=LPGA;
   SYMBOL V=DOT;
   PLOT FAIRWAYS*GREENS;
RUN;
```

```
data one;
  mean1=0;  *mean for y1;
  mean2=10; *mean for y2;
  sig1=2;   *SD for y1;
  sig2=5;   *SD for y2;
  rho=0.5;  *Correlation between y1 and y2;
  do i = 1 to 1000;
     r1 = rannor(1245);
     r2 = rannor(2923);
     y1 = mean1 + sig1*r1;
     y2 = mean2 + rho*sig2*r1+sqrt(sig2**2-sig2**2*rho**2)*r2;
     output;
  end;
keep y1 y2;
*proc print;
run;
```

Matrix from and into SAS data,

▪프로그램

```
DATA ONE;
   INPUT Y X1 X2;
   CARDS;
1  1  3
1  2  3
2  3  4
2  4  5
3  5  5
RUN;

PROC IML;
   USE ONE;
   /*SAS 데이터 이용 행렬 만들기 */
   READ ALL VAR{X1 x2} INTO X;
   READ ALL VAR{Y} INTO Y;
   X1=J(NROW(Y),1)||X;
   PRINT X, Y, X1;
   BHAT=INV(X1`*X1)*X1`*Y;
   YHAT=X1*BHAT;
   PRINT BHAT, YHAT;
   RES=Y-YHAT;
   DS=Y||X||DS||YHAT||RES;
   PRINT DS;
   /*행렬 데이터 만들기*/
   CREATE ONE2 VAR{Y X1 X2 YHAT RES};
   APPEND FROM DS;
RUN;

PROC PRINT DATA=ONE2;
RUN;
```

▪결과

▪ matrix

```
      X                  Y              X1
      1      3           1      1       1      3
      2      3           1      1       2      3
      3      4           2      1       3      4
      4      5           2      1       4      5
      5      5                  1       5      5

     BHAT                         DS
       0.3
       0.5
 -4.44E-15     1      1      3     0.8     0.2
     YHAT      1      2      3     1.3    -0.3
               2      3      4     1.8     0.2
       0.8     2      4      5     2.3    -0.3
       1.3     3      5      5     2.8     0.2
       1.8
       2.3
       2.8
```

▪ SAS 데이터 "one2"

OBS	Y	X1	X2	YHAT	RES
1	1	1	3	0.8	0.2
2	1	2	3	1.3	-0.3
3	2	3	4	1.8	0.2
4	2	4	5	2.3	-0.3
5	3	5	5	2.8	0.2