

STAT 430

SAS Examples SAS5

=====

ssh xyz@glue.umd.edu , tap sas913 (old sas82), sas  
https://www.statlab.umd.edu/sasdoc/sashtml/onldoc.htm

CH5: CORR & SIMPLE LINEAR REFRESSION

=====

1. Pearson Correlation
2. Spearman Correlation
3. Partial Correlation
4. Simple Linear Regression
5. Adding a Quadratic Term
6. Transforming Data to Get a Better Fit
7. Computing Within-Subject Slopes
8. An Additional Example.

1. Pearson Correlation

-----

OPTION PS=35 LS=70;

- a. From PROC CORR get also simple statistics.
- b. CORR ignores missing data.

DATA SET1;

INPUT GENDER \$ HEIGHT WEIGHT AGE;

DATALINES;

M 68 155 23

F 61 99 20

F 63 115 21

M 70 205 45

M 69 170 .

F 65 125 30

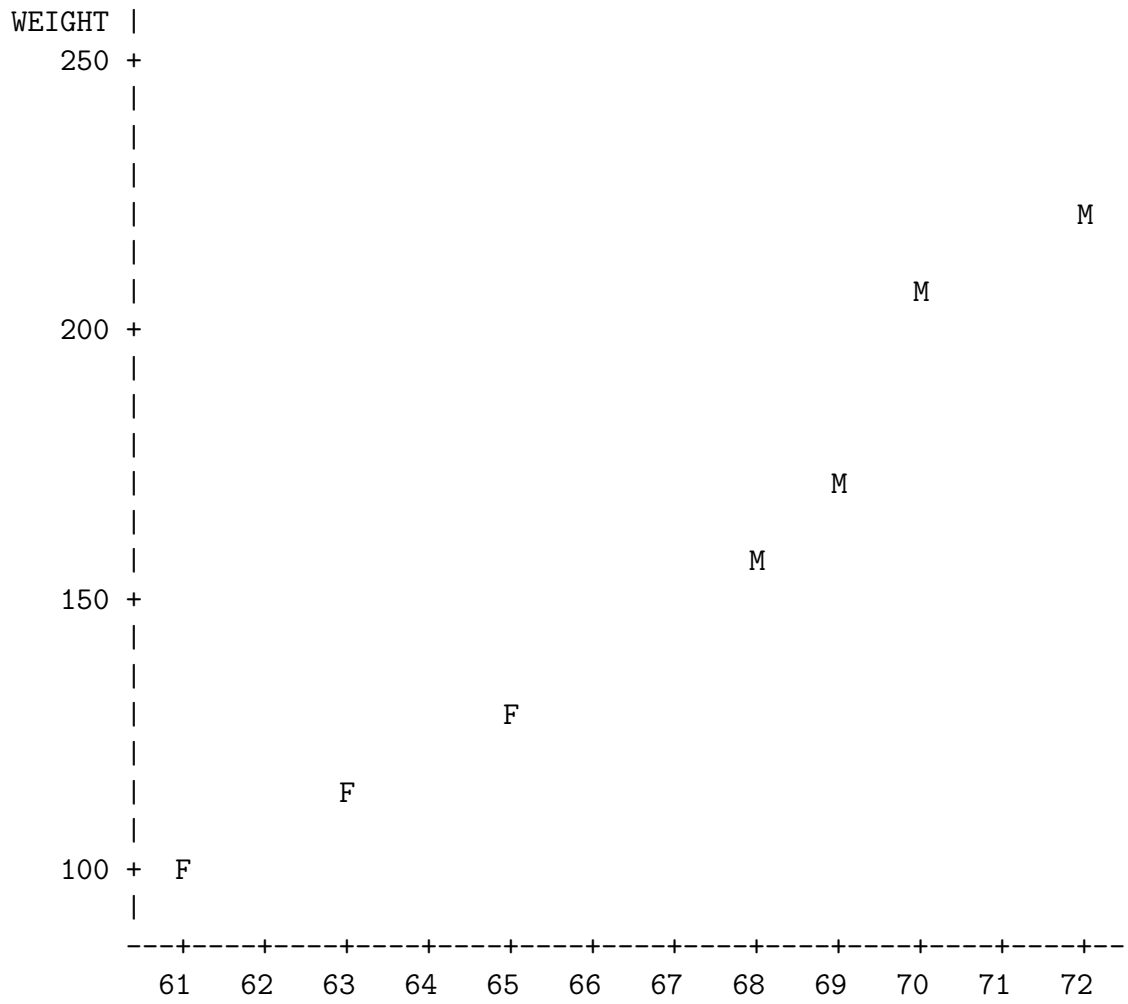
M 72 220 48

;

Good idea to first plot to get an idea of the relationships!!!

```
PROC PLOT DATA=SET1;  
PLOT WEIGHT*HEIGHT=GENDER;  
RUN;
```

Plot of WEIGHT\*HEIGHT. Symbol is value of GENDER.

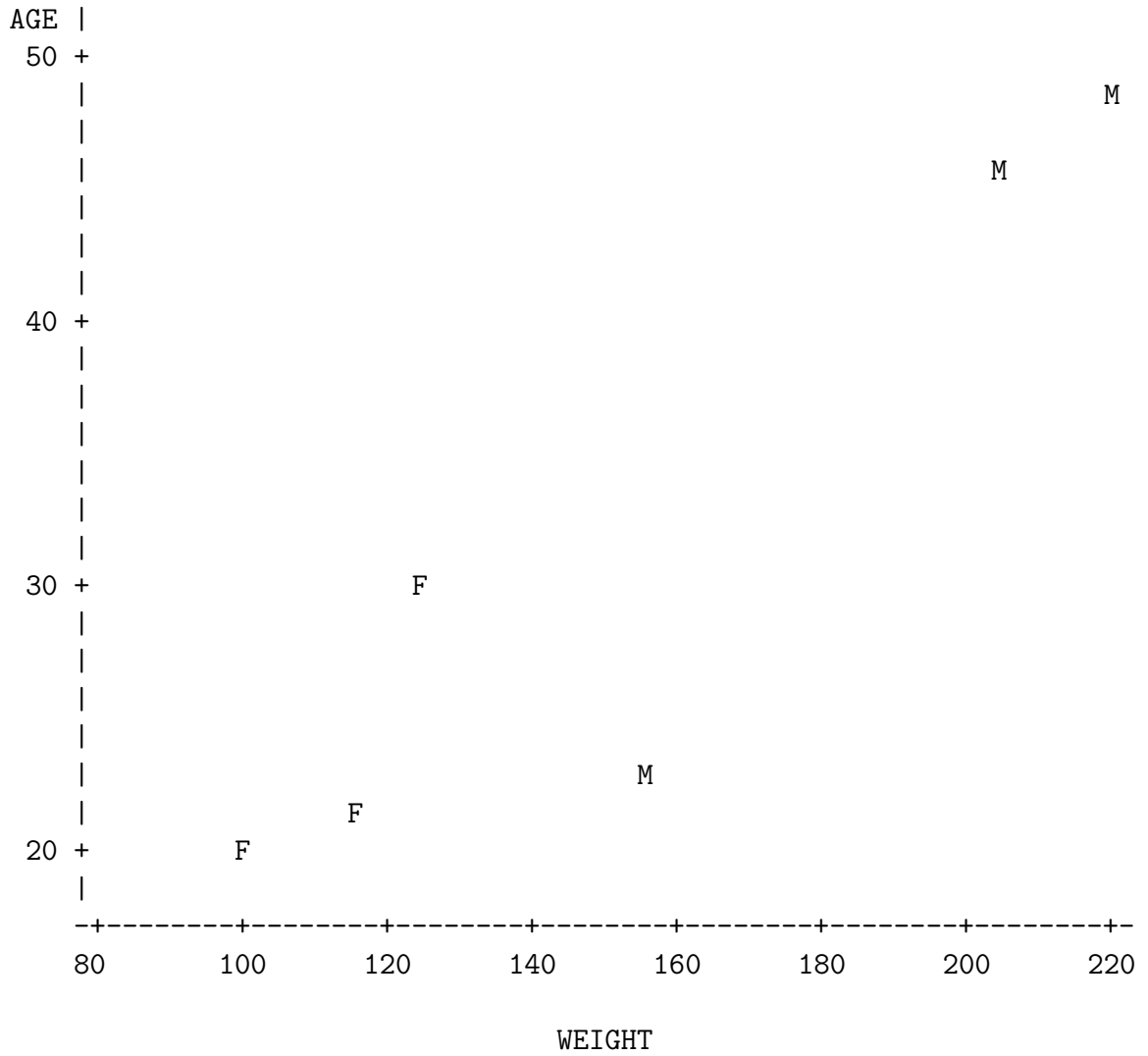


HEIGHT

Expect high corr between HEIGHT and WEIGHT!!!

```
PROC PLOT DATA=SET1;  
PLOT AGE*WEIGHT=GENDER;  
RUN;
```

Plot of AGE\*WEIGHT. Symbol is value of GENDER.



NOTE: 1 obs had missing values.

Expect a pretty high corr between WEIGHT and AGE!!!

```
PROC CORR DATA=SET1;  
TITLE 'EXAMPLE OF CORR MATRIX';  
VAR HEIGHT WEIGHT AGE;  
RUN;
```

3 Variables: HEIGHT WEIGHT AGE

Simple Statistics

Variable	N	Mean	Std Dev	Sum
HEIGHT	7	66.85714	3.97612	468.00000
WEIGHT	7	155.57143	45.79613	1089
AGE	6	31.16667	12.41639	187.00000

Simple Statistics

Variable	Minimum	Maximum
HEIGHT	61.00000	72.00000
WEIGHT	99.00000	220.00000
AGE	20.00000	48.00000

EXAMPLE OF CORR MATRIX

2

The CORR Procedure

Pearson Correlation Coefficients  
 Prob > |r| under H0: Rho=0  
 Number of Observations

	HEIGHT	WEIGHT	AGE
HEIGHT	1.00000	0.97165	0.86614
		0.0003	0.0257
	7	7	6
WEIGHT	0.97165	1.00000	0.92496
	0.0003		0.0082
	7	7	6
AGE	0.86614	0.92496	1.00000
	0.0257	0.0082	
	6	6	6

NOTE: The corr between HEIGHT, WEIGHT and AGE was obtained from the 6 full pairs excluding the row "M 69 170 ." where AGE is missing.

If only need corr(AGE,HEIGHT) AND corr(AGE,WEIGHT) use WITH!!!

```
PROC CORR DATA=SET1 PEARSON NOSIMPLE;
VAR HEIGHT WEIGHT;
WITH AGE;
RUN;
```

The CORR Procedure

```
1 With Variables:    AGE
2   Variables:     HEIGHT  WEIGHT
```

Pearson Correlation Coefficients  
 Prob > |r| under H0: Rho=0  
 Number of Observations

	HEIGHT	WEIGHT
AGE	0.86614	0.92496
	0.0257	0.0082
	6	6

## 2. Spearman Correlation

-----

- c. Use the option NOSIMPLE if simple statistics are not needed.
- d. To get SPEARMAN corr, we must use the PEARSON option as well.
- e. In our exmple we have:

Rank of Height1 is 4, Rank of Height2 is 1, Rank of Height3 is 2 etc.

Ranks(Height)= 4 1 2 6 5 3 7

Ranks(Weight)= 4 1 2 6 5 3 7

Ranks(Age)= 3 1 2 5 . 4 6

NOTE:                   4 1 2 6 5 3 7                   4 1 2 6 5 3 7  
 Weight 155 99 115 205 170 125 220   Height 68 61 63 70 69 65 72

\*\*\*SPEARMAN is the correlation of the ranks. Since in this example  
 Weight and Height have the same ranks, SC=1.

```
PROC CORR DATA=SET1 PEARSON SPEARMAN NOSIMPLE;
TITLE 'EXAMPLE OF CORR MATRIX';
VAR HEIGHT WEIGHT AGE;
RUN;
```

The CORR Procedure

Spearman Correlation Coefficients

Prob > |r| under H0: Rho=0

Number of Observations

	HEIGHT	WEIGHT	AGE
HEIGHT	1.00000	1.00000	0.94286
		<.0001	0.0048
	7	7	6
WEIGHT	1.00000	1.00000	0.94286
	<.0001		0.0048
	7	7	6
AGE	0.94286	0.94286	1.00000
	0.0048	0.0048	
	6	6	6

3. Partial Correlation

-----

We can remove the effect of AGE using Partial Cor:

```
PROC CORR DATA=SET1 PEARSON NOSIMPLE;  
VAR HEIGHT WEIGHT;  
PARTIAL AGE;  
RUN;
```

The CORR Procedure

```

1 Partial Variables:  AGE
2      Variables:    HEIGHT  WEIGHT

```

Pearson Partial Correlation Coefficients, N = 6  
 Prob > |r| under H0: Partial Rho=0

	HEIGHT	WEIGHT
HEIGHT	1.00000	0.91934 0.0272
WEIGHT	0.91934 0.0272	1.00000

#### 4. Simple Linear Regression

-----

Use: PROC REG < options > ;

FIT:  $y=a+bx$

-----

General form of PROC REG is:

```

PROC REG DATA=DATASET;
MODEL DEPENDENT VARIABLE(S) = INDEPENDENT VARIABLES / OPTIONS;
RUN;

```

```

OPTION PS=35 LS=70;

```

```

DATA SET1;
INPUT GENDER $ HEIGHT WEIGHT AGE;
DATALINES;
M 68 155 23

```



```

F 61 99 20
F 63 115 21
M 70 205 45
M 69 170 .
F 65 125 30
M 72 220 48
;

```

```

PROC REG DATA=SET1;
MODEL WEIGHT=HEIGHT; (Weight = a + b*Height + e)
RUN;

```

The REG Procedure  
 Model: MODEL1  
 Dependent Variable: WEIGHT

n=7,r=2,q=1,n-r=5,n-1=6,k=1 (one slope b), H<sub>0</sub>: b=0

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	q=1	11880	11880	84.45	0.0003
Error	n-r=5	703.38705	140.67741		
Corrected Total	n-1=6	12584			

Root MSE	11.86075	R-Square	0.9441
Dependent Mean	155.57143	Adj R-Sq	0.9329
Coeff Var	7.62399		

Note:

$\text{Coeff Var} = (\text{Root MSE}) / (\text{Dependent Mean}) * 100 =$   
 $11.86075 * 100 / 155.57143 = 7.62399$

Note:  $\text{Adj R-Sq} = \frac{\{(n-1)R^2 - k\}}{(n-k-1)} = \frac{(6 \cdot 0.9441 - 1)}{(7-1-1)} = 0.93292$

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	-592.64458	81.54217	-7.27	0.0008
HEIGHT	1	11.19127	1.21780	9.19	0.0003

Note:

$P(|T| > |-7.27|) = 2 \cdot P(T < -7.27) = 0.000769$  with  $df = n-r = 5$ .

$P(|T| > 9.19) = 2 \cdot P(T < -9.19) = 0.000256$  with  $df = n-r = 5$ .

-----  
Fitted model:  $\text{WEIGHT} = -592.64458 + 11.19127 \cdot \text{HEIGHT}$   
-----

Plotting predicted and actual observation on the same plot:  
-----

```
PROC REG DATA=SET1;  
MODEL WEIGHT=HEIGHT;  
PLOT PREDICTED. *HEIGHT = 'P' WEIGHT*HEIGHT='O' / OVERLAY;  
RUN;
```

The book suggests that but we get a graphical window we cannot select and paste. So, to get around this problem, define the regression line in the DATA statement and then plot!

```
DATA SET1;  
INPUT GENDER $ HEIGHT WEIGHT AGE;  
YHAT = -592.64458 + 11.19127*HEIGHT; <---Estimated regression line  
DATALINES;  
M 68 155 23  
F 61 99 20
```

```

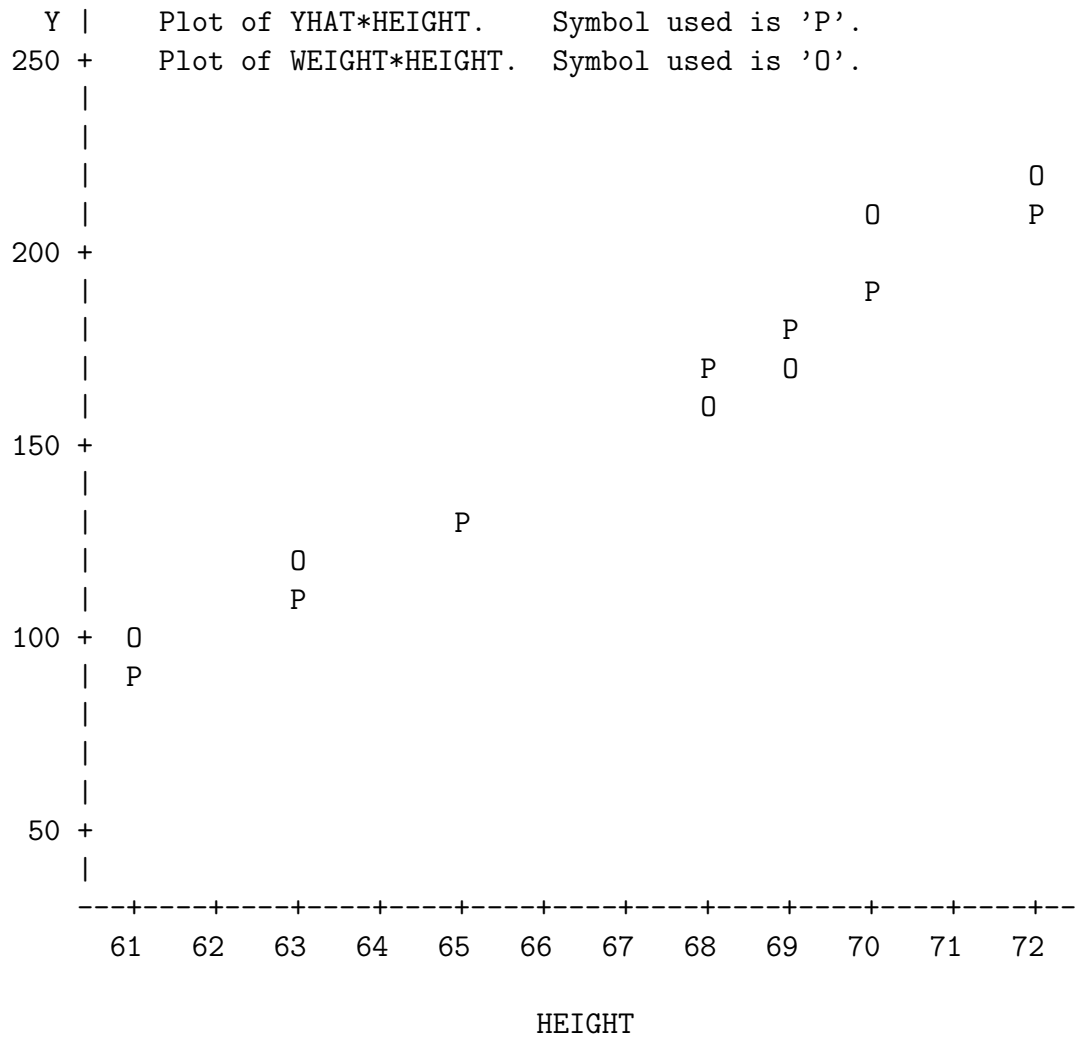
F 63 115 21
M 70 205 45
M 69 170 .
F 65 125 30
M 72 220 48
;

```

```

PROC PLOT DATA=SET1;
PLOT YHAT*HEIGHT='P' WEIGHT*HEIGHT='O'/OVERLAY;
RUN;

```



NOTE: 1 obs hidden. <----- Obs too close to Pred

Plot of residuals once using RESIDUAL. and once  
using the same trick of defining the residuals  
in the DATA statement:

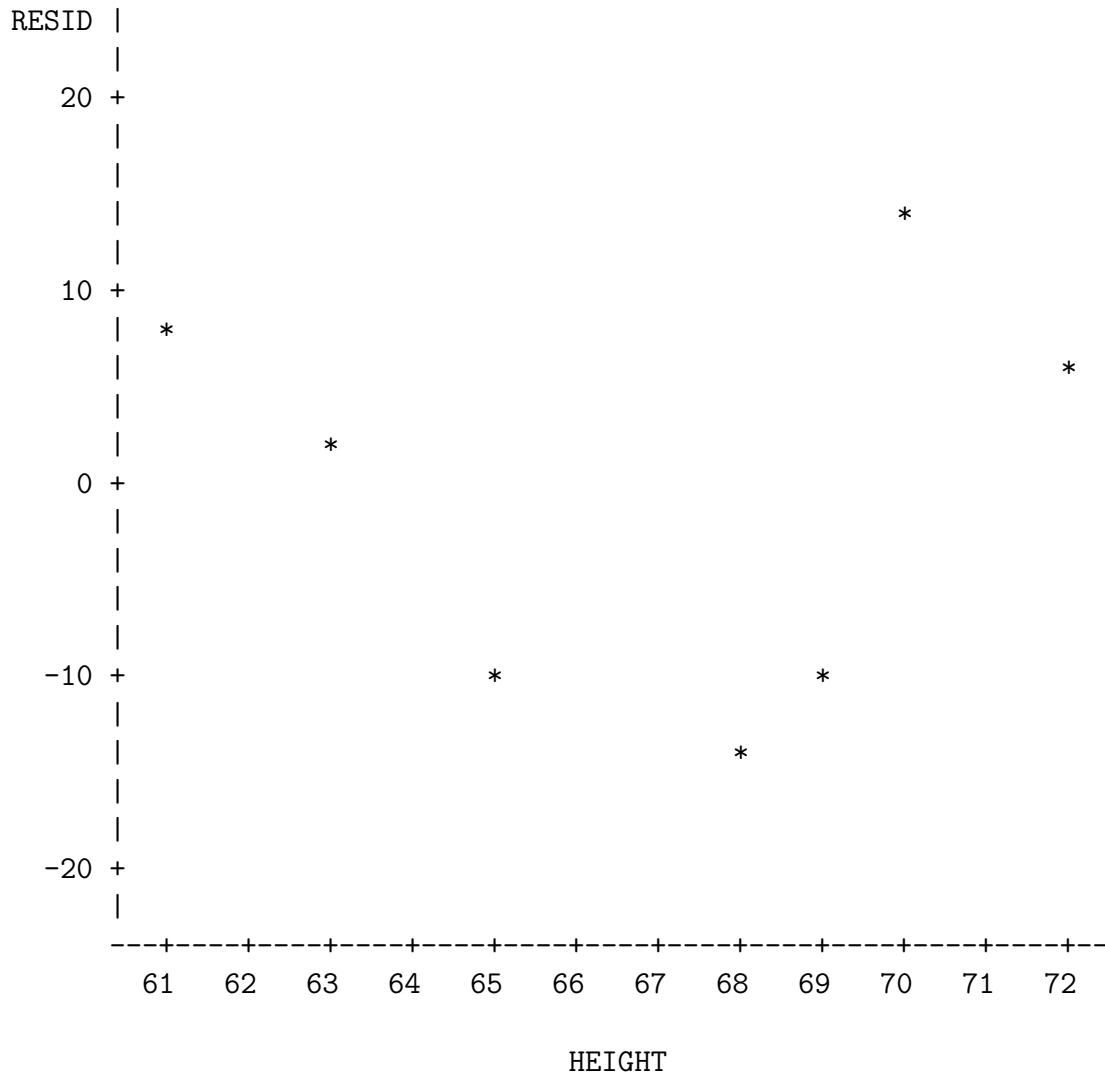
```
DATA SET1;
INPUT GENDER $ HEIGHT WEIGHT AGE;
/* My definition of YHAT and RESID for resid plot not using RESIDUAL.*/
YHAT = -592.64458 + 11.19127*HEIGHT; <---Estimated regression line
RESID=WEIGHT-YHAT;
DATALINES;
M 68 155 23
F 61 99 20
F 63 115 21
M 70 205 45
M 69 170 .
F 65 125 30
M 72 220 48
;
```

```
PROC REG DATA=SET1;          <--- Does regression and gives
MODEL WEIGHT=HEIGHT;         Residual plot using RESIDUAL.
PLOT RESIDUAL. *HEIGHT='0';  But uses GUI of SAS!!!
RUN;
```

To bypass GUI of SAS, we can declare the new variables  
YHAT = -592.64458 + 11.19127\*HEIGHT;  
RESID=WEIGHT-YHAT;  
and then use:

```
PROC PLOT DATA=SET1;
PLOT RESID*HEIGHT='*';
RUN;
```

Plot of RESID\*HEIGHT. Symbol used is '\*'.



5. Adding a quadratic term: Get higher  $R^2$ !!!

---

The residual plot suggests we add a quadratic term in the regression eq!!!

```

DATA SET1;
INPUT GENDER $ HEIGHT WEIGHT AGE;
/* My definition of YHAT and RESID for resid plot not using RESIDUAL.*/
YHAT = 2321.12131 + -76.84468*HEIGHT + 0.66290*HEIGHT**2;
RESID=WEIGHT-YHAT;
HEIGHT2=HEIGHT**2;
DATALINES;
M 68 155 23
F 61 99 20
F 63 115 21
M 70 205 45
M 69 170 .
F 65 125 30
M 72 220 48
;

PROC REG DATA=SET1;           Model:  $w=a+b_1*H+b_2*H^2+e$ 
MODEL WEIGHT=HEIGHT HEIGHT2; <--- Does regression and gives
PLOT RESIDUAL. *HEIGHT='0';   Residual plot. But plot is on gui!
RUN;

```

The REG Procedure, Model: MODEL1, Dependent Variable: WEIGHT  
 Analysis of Variance:  $n=7, r=3, q=2, k=2$  ( $b_1, b_2$ ),  $H_0: b_1=b_2=0$

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	q=2	12261	6130.45691	75.97	0.0007
Error	n-r=4	322.80046	80.70012		
Corrected Total	n-1=6	12584			

Root MSE	8.98332	R-Square	0.9743<--Higher!
Dependent Mean	155.57143	Adj R-Sq	0.9615
Coeff Var	5.77440		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	2321.12131	1343.15025	1.73	0.1590
HEIGHT	1	-76.84468	40.54924	-1.90	0.1310
HEIGHT2	1	0.66290	0.30525	2.17	0.0956

Note: df=4

$2 * P(T_4 < -1.73) = 0.1590$ , but  $2 * P(T_1 < -1.73) = 0.337$

$2 * P(T_4 < -1.90) = 0.1302$

$2 * P(T_4 < -2.17) = 0.0958$

My trick: To bypass GUI, define YHAT and RESID!!! in DATA statement as above:

$YHAT = 2321.12131 + -76.84468 * HEIGHT + 0.66290 * HEIGHT ** 2;$

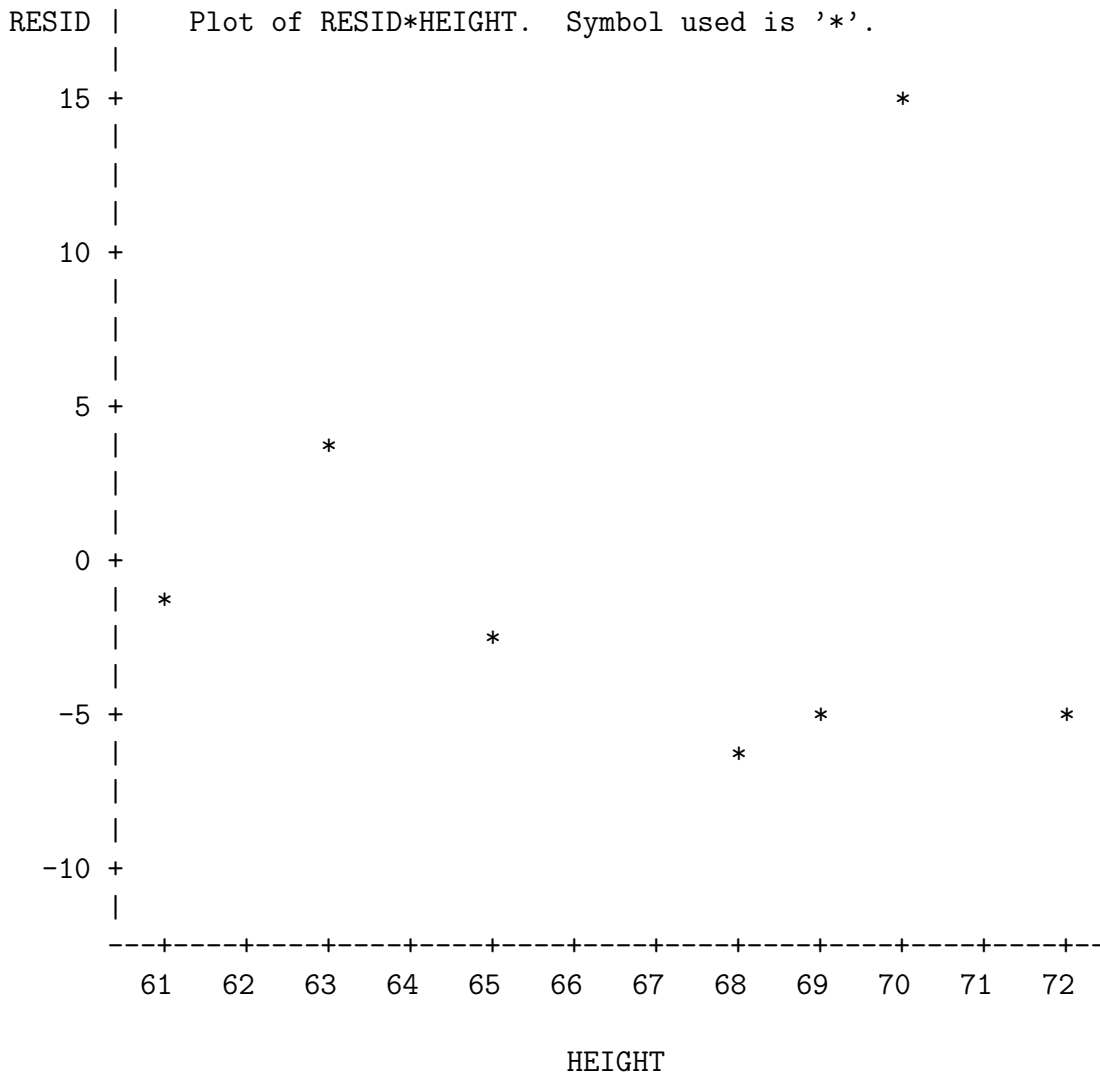
$RESID = WEIGHT - YHAT;$

Then use:

```
PROC PLOT DATA=SET1;
```

```
PLOT RESID*HEIGHT='*';
```

```
RUN;
```



## 6. Transforming Data to get a better fit

-----  
 OPTION PS=35 LS=70;

DATA HEART;  
 INPUT DOSE HR;  
 DATALINES;  
 2 60





DOSE

NOTE: 1 obs hidden. <-- SINCE WE HAVE TWO PAIRS (16,70).

The plot is not exactly linear, but appears on log scale. So first we run linear simple regression and then we'll use log(DOSE).

```
PROC REG DATA=HEART;  
MODEL HR=DOSE;  
RUN;
```

The REG Procedure, Model: MODEL1, Dependent Variable: HR  
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	233.48441	233.48441	49.01	0.0001
Error	8	38.11559	4.76445		
Corrected Total	9	271.60000			

Root MSE	2.18276	R-Square	0.8597	<--Can be improved.	
Dependent Mean	66.20000	Adj R-Sq	0.8421		
Coeff Var	3.29722				

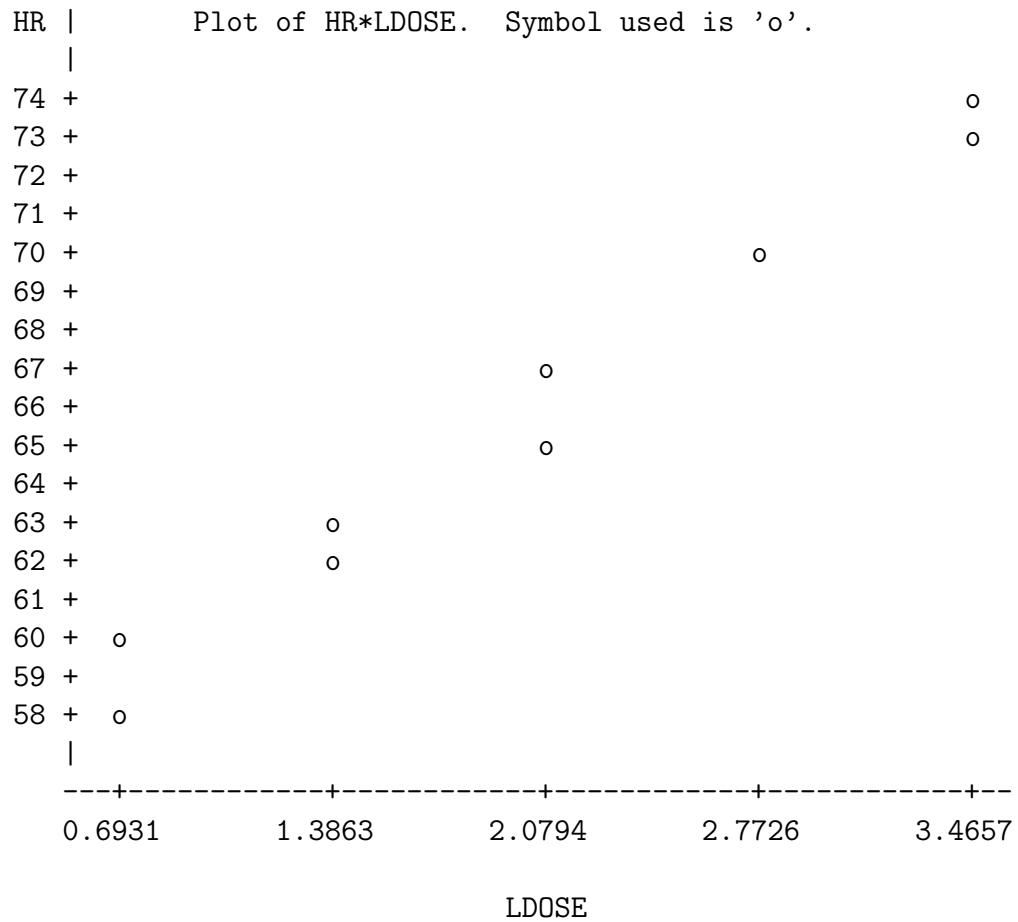
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	60.70833	1.04492	58.10	<.0001
DOSE	1	0.44288	0.06326	7.00	0.0001

NOW USE THE LOG TRANSFORMATION!!!

```
DATA HEART;
INPUT DOSE HR;
LDOSE=LOG(DOSE); <--Dose on log-scale.
DATALINES;
  2 60
  2 58
  4 63
  4 62
  8 67
  8 65
 16 70
 16 70
 32 74
 32 73
;

PROC PLOT DATA=HEART;
PLOT HR*LDOSE='o';
RUN;
```



NOTE: 1 obs hidden.

LOOKS LINEAR!!!

```
PROC REG DATA=HEART;
MODEL HR=LDOSE; <---HR=a+b*log(DOSE)+e
RUN;
```

Get a much larger  $R^2=0.9810$ . Before we had  $R^2=0.8597$ .

The REG Procedure  
 Model: MODEL1  
 Dependent Variable: HR

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	266.45000	266.45000	413.90	<.0001
Error	8	5.15000	0.64375		
Corrected Total	9	271.60000			

Root MSE	0.80234	R-Square	0.9810
Dependent Mean	66.20000	Adj R-Sq	0.9787
Coeff Var	1.21199		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	55.25000	0.59503	92.85	<.0001
LDOSE	1	5.26584	0.25883	20.34	<.0001

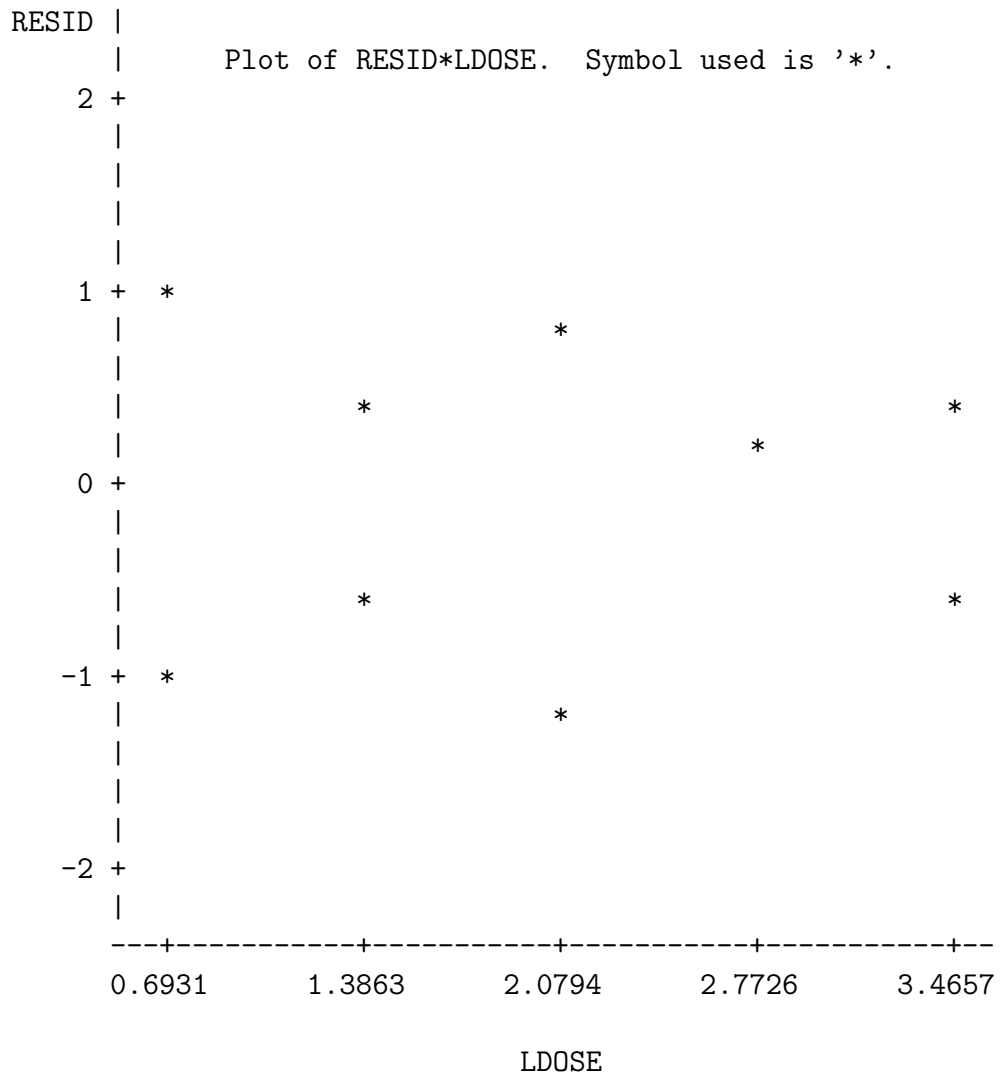
```

PROC REG DATA=HEART;
MODEL HR=LDOSE;          <--- Does regression and gives
PLOT RESIDUAL. *LDOSE='0';      Residual plot. But plot is on gui!!!
RUN;
```

Simpler: Define the residuals in DATA STATEMENT

```
DATA HEART;
INPUT DOSE HR;
LDOSE=LOG(DOSE);
/* My definition of YHAT and RESID for resid plot not using RESIDUAL.*/
YHAT =55.25000 + 5.26584*LDOSE;
RESID=HR-YHAT;
DATALINES;
  2 60
  2 58
  4 63
  4 62
  8 67
  8 65
 16 70
 16 70
 32 74
 32 73
;

PROC PLOT DATA=HEART;
PLOT RESID*LDOSE='*';
RUN;
```



NOTE: 1 obs hidden.

## 7. Computing Within-Subject Slopes

Get slope from repeated measures for each subject. This is done

with the option OUTSET in REG which creates a new data set listing the slopes and intercepts.

```
OPTION PS=35 LS=70;
```

```
DATA TEST;  
INPUT ID GROUP $ TIME SCORE;  
DATALINES;  
1 A 1 2  
1 A 2 5  
1 A 3 7  
2 A 1 4  
2 A 2 6  
2 A 3 9  
3 B 1 8  
3 B 2 6  
3 B 3 2  
4 B 1 8  
4 B 2 7  
4 B 3 3  
;
```

```
PROC SORT DATA=TEST; <--Need this first  
BY ID;  
RUN;
```

```
PROC REG OUTSET=SLOPES DATA=TEST;  
BY ID;  
ID GROUP;  
MODEL SCORE=TIME / NOPRINT;  
RUN;
```

```
PROC PRINT DATA=SLOPES;  
RUN;
```



Obs	ID	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	Intercept	TIME	SCORE
1	1	MODEL1	PARMS	SCORE	0.40825	-0.3333	2.5	-1
2	2	MODEL1	PARMS	SCORE	0.40825	1.3333	2.5	-1
3	3	MODEL1	PARMS	SCORE	0.81650	11.3333	-3.0	-1
4	4	MODEL1	PARMS	SCORE	1.22474	11.0000	-2.5	-1

We can also compare the slope means between groups by a t-test:

```
PROC TTEST DATA=SLOPES;
CLASS GROUP;
VAR TIME;
RUN;
```

-----

Another Example: Blood Pressure

=====

```
OPTION PS=35 LS=70;
```

```
DATA BP;
INPUT SysBP DiasBP;
DATALINES;
180 110
190 108
178 100
170 100
180 98
168 88
160 80
160 80
172 86
```

```
170 86
140 80
130 72
128 70
;
```

```
PROC CORR DATA=BP PEARSON NOSIMPLE;
VAR SysBP DiasBP;
RUN;
```

### The SAS System

#### The CORR Procedure

2 Variables: SysBP DiasBP

Pearson Correlation Coefficients, N = 13  
Prob > |r| under H0: Rho=0

	SysBP	DiasBP
SysBP	1.00000	0.88221 <.0001
DiasBP	0.88221 <.0001	1.00000

```
PROC REG DATA=BP;
MODEL SysBP=DiasBP;
RUN;
```

The REG Procedure  
 Model: MODEL1  
 Dependent Variable: SysBP

Number of Observations Read 13  
 Number of Observations Used 13

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3574.85010	3574.85010	38.61	<.0001
Error	11	1018.38066	92.58006		
Corrected Total	12	4593.23077			

Root MSE 9.62185 R-Square 0.7783  
 Dependent Mean 163.53846 Adj R-Sq 0.7581  
 Coeff Var 5.88354

Note:  $R^2=0.7783^2=0.88221 = \text{Corr}(\text{SysBP}, \text{DiasBP})= 0.88221$

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	45.53172	19.17710	2.37	0.0369
DiasBP	1	1.32477	0.21319	6.21	<.0001