

PubH 7405: REGRESSION ANALYSIS



SLR: REVIEWS & COMPUTATIONS

CORRELATION & REGRESSION

- We have 2 continuous measurements made on each subject, one is the response variable Y, the other predictor X. There are two types of analyses:
- Correlation: is concerned with the association between them, measuring the strength of the relationship; the aim is to determine if they are correlated – the roles are exchangeable.
- Regression: To predict response from predictor.

You normally like to proceed to performing prediction **only if** the association is strong enough. However, in practice, “correlation analysis” only covers association whereas “regression analysis” would cover **both** association and prediction simultaneously.

We start with a statistical/algebraic model, called “Normal Error Regression Model” which can be easily extended into a multivariate model for use in the cases we have more than one predictors. Some of the results can be used to for a Geometric Model/Representation.

(Simple)

Normal Error Regression Model :

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

$$\varepsilon \in N(0, \sigma^2)$$

The Mean Response :

$$\mathbf{E}(Y) = \beta_0 + \beta_1 \mathbf{x}$$

(Multiple)

Normal Error Regression Model :

$$Y = \beta_0 + \beta_1 \mathbf{x}_1 + \cdots + \beta_k \mathbf{x}_k + \varepsilon$$

$$\varepsilon \in N(0, \sigma^2)$$

The Mean Response :

$$\mathbf{E}(Y) = \beta_0 + \beta_1 \mathbf{x}_1 + \cdots + \beta_k \mathbf{x}_k + \varepsilon$$

Main Results come as four sets of theorems:

Theorems 1A and 1B: About the **Slope**

Theorems 2A and 2B: About the **Intercept**

Theorems 3A and 3B: About the **Mean Response**

Theorems 4A and 4B: About **Individual Response**

The first theorem of each set specifies the sampling distribution (Normal), the second lead to a t-distribution - same t-distribution with **df = n-2**

Plus Theorem 5 on the sampling distribution of MSE/σ^2 ; **Together, the five theorems allow us to draw inferences, in the form of Confidence Intervals – as well as to test for Independence – two t-tests and an F test (with identical results).**

FITTED VALUE & RESIDUAL

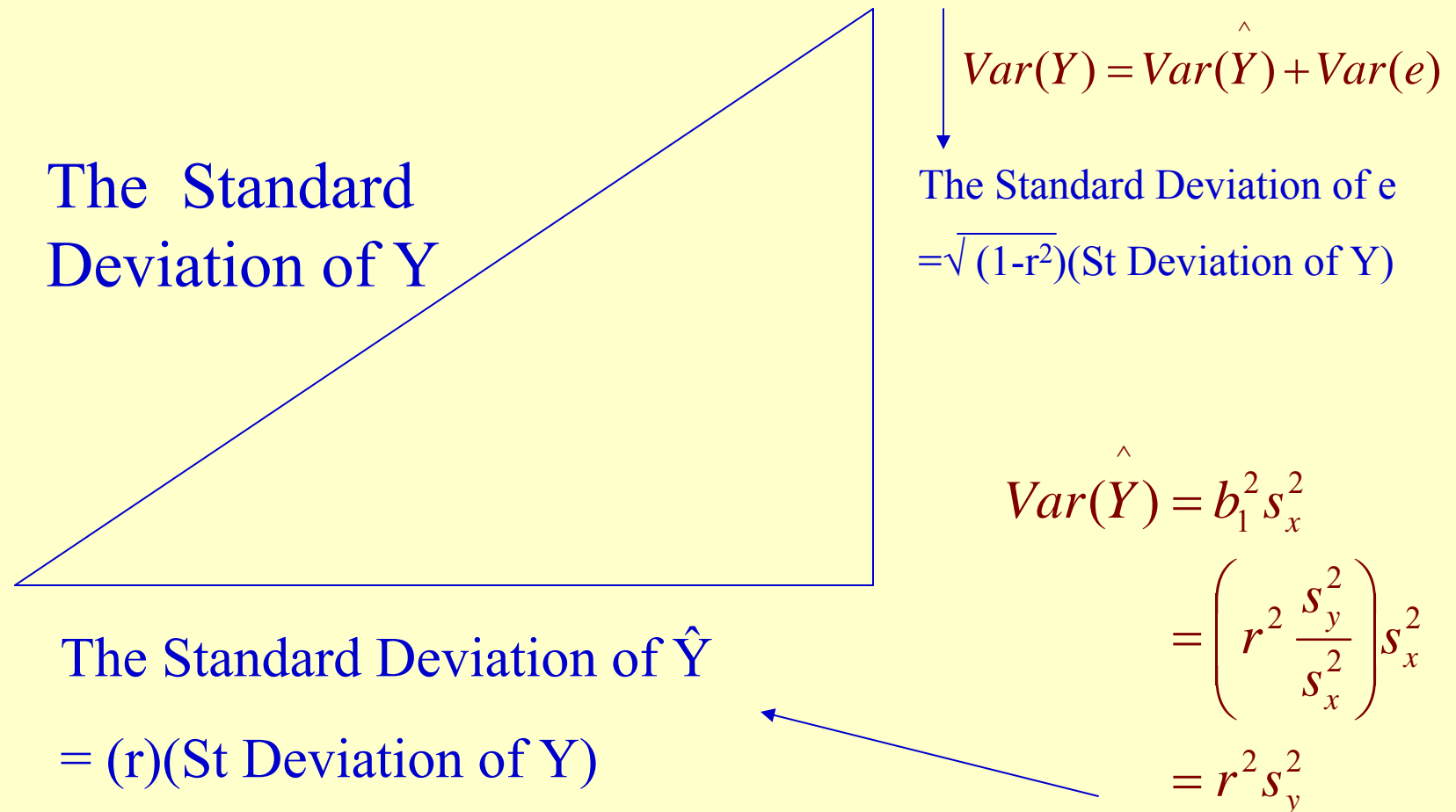
$$e = Y - \hat{Y}$$

$$Y = \hat{Y} + e$$

$$\begin{aligned} \text{Var}(Y) &= \text{Var}(\hat{Y}) + \text{Var}(e) - 2\text{Cov}(\hat{Y}, e) \\ &= \text{Var}(\hat{Y}) + \text{Var}(e) \end{aligned}$$

Here the regression line is used to predict Y from X; the predicted/fitted value is \hat{Y} , and the error/residual of this prediction is e. In the above formula, (1) $\text{Var}(Y)$ is the total variance of interest, (2) $\text{Var}(\hat{Y})$ is the “explained” variance, and (3) $\text{Var}(e)$ is the “unexplained” variance.

Pythagorean Representation



$$e = Y - \hat{Y}$$

$$Y = \hat{Y} + e$$

$$\text{Var}(Y) = \text{Var}(\hat{Y}) + \text{Var}(e) + 2\text{Cov}(\hat{Y}, e)$$

$$= \text{Var}(\hat{Y}) + \text{Var}(e)$$

$$\text{Var}(e) = \text{Var}(Y) - \text{Var}(\hat{Y})$$

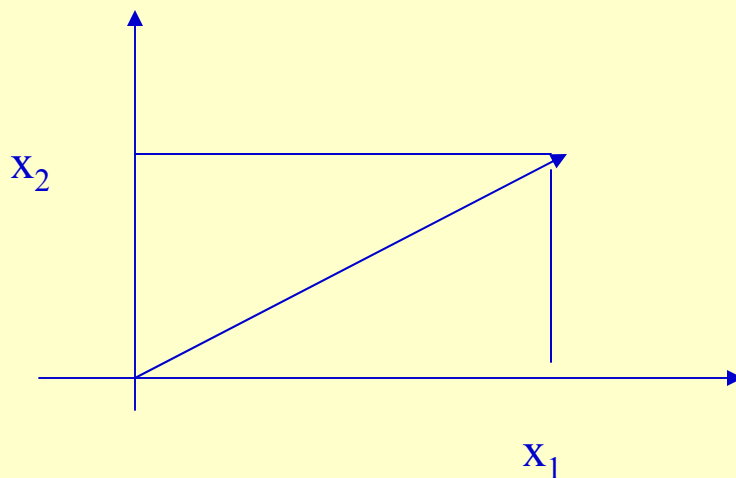
$$= s_y^2 - r^2 s_y^2$$

$$= (1 - r^2) s_y^2$$

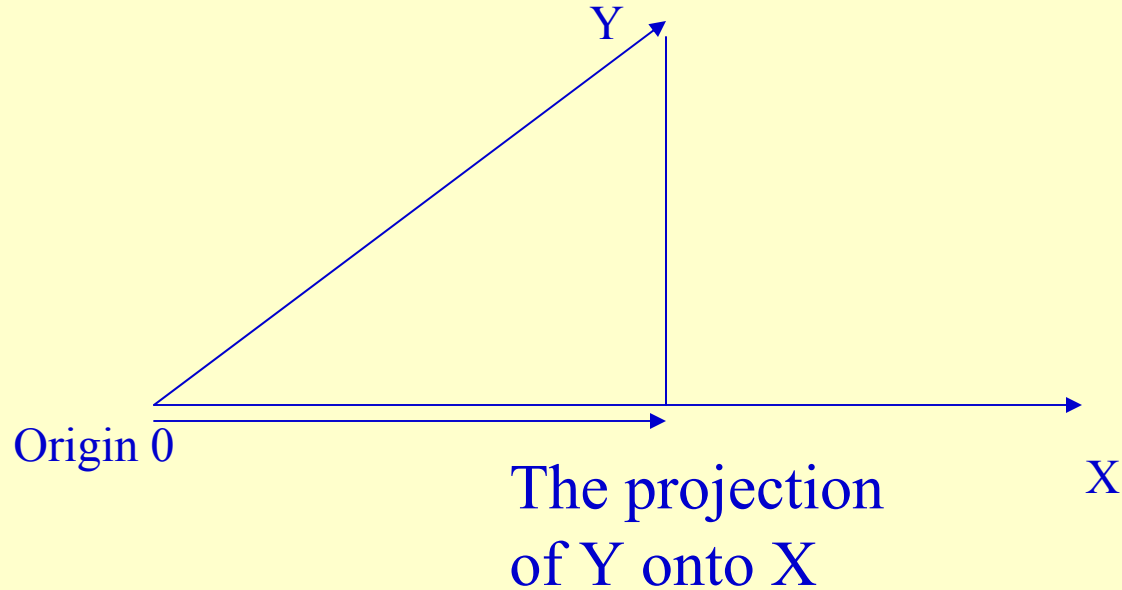
Result is :

$$r^2 \leq 1; \text{ or } : -1 \leq r \leq 1$$

Now think of representing the “regression data” not as “n pairs/sets of numbers” but by two points X and Y in an n -dimensional space. Think of each point as a “vector”, an “arrow” drawn from the origin to the point. The following picture is a two-dimensional vector:



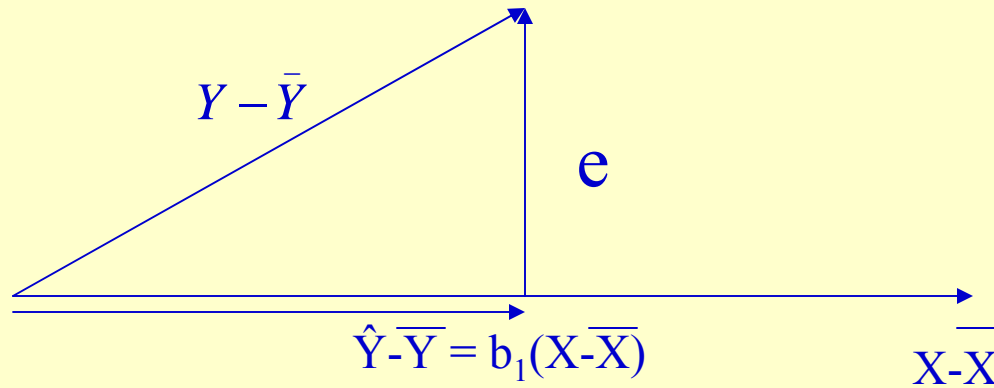
Now, consider 2 vectors, X and Y , in the same n -dimensional space. The “projection” of Y on X is obtained by dropping a perpendicular from Y onto X .



What we get is the multiple of X closest to Y

A GEOMETRY MODEL

- Regression involves two points in the n -dimensional space
- One point represents the deviation from X to its average \bar{X} ,
- The other point represents the deviation of Y from its average \bar{Y} .
- The “Regression of Y on X ” can be seen as the projection of the vector $(Y - \bar{Y})$ on the vector $(X - \bar{X})$.



The vector residuals “e” is perpendicular to vector $(X - \bar{X})$:

$$\text{Cov}(e, X) = \frac{1}{n} \sum_{i=1}^n e_i (x_i - \bar{x}) = 0$$

Besides other applications, this “Geometry Model” is a different way to express the previous “Analysis of Variance” because, in the n-dimensional space, the squares of the lengths are the variances and:

- (1) The standard deviation of Y is the length of $(Y - \bar{Y})$,
- (2) The standard deviation of \hat{y} is the length of the projection $(\hat{Y} - \bar{Y})$, and
- (3) The standard deviation of e is the length of the residual vector e

PROTOTYPE EXAMPLE

Age (x)	SBP (y)
42	130
46	115
42	148
71	100
80	156
74	162
70	151
80	156
85	162
72	158
64	155
81	160
41	125
61	150
75	165

Will use for Illustration

options ls=79; BASIC DATA DESCRIPTION

```
title "SBP versus Age";
```

```
data SBP;
```

```
input age pressure;
```

```
cards;
```

```
42 130
```

```
46 115
```

```
42 148
```

```
71 100
```

```
80 156
```

```
74 162
```

```
70 151
```

```
80 156
```

```
85 162
```

```
72 158
```

```
64 155
```

```
81 160
```

```
41 125
```

```
61 150
```

```
75 165
```

```
;
```

Notes:

- (1) Can use “**data lines**” instead of “**cards**”
- (2) Good enough for smaller data sets
- (3) For a larger data set, save it as “abc.dat” or “abc.xls” and refer to it or import it; use **PROC IMPORT** (a bit later).

Same order as in the data

DESCRIPTIVE STATISTICS

```
options ls=79;
title "Descriptive Statistics for SBP
versus Age";
data SBP;
input X Y;
  label X = 'Age'
        Y = 'Blood Pressure';

cards;
42 130
46 115
...
75 165
;
proc PRINT data=SBP;
Var X Y;
run;
proc UNIVARIATE data=SBP;
run;
```

PRINT helps to check for typos

UNIVARIATE provides typical data summaries such as mean, range, standard deviation, etc...

From TA's HANDOUT

File name

```
Proc IMPORT out=work.hw1
datafile="C:\Documents and Settings\ADCS-C381Mayo-User\Desktop\CH01PR19.xls"
DBMS=EXCEL2000 REPLACE;
GETNAMES=YES;
run;
data hw1;
set work.hw1;
run;
```

Important Part:

Showing HOW to read in
data file (its name & location)

```
Proc MEANS data=hw1 STDERR maxdec=1;
```

```
Var x;
```

```
run;
```

Specify max # of decimal places

Request Standard Error of the Mean

```
Proc print data=hw1(obs=20) noobs;
```

```
run;
```

Suppress the observation number

MORE OPTIONS for Proc Univariate

```
options ls=79;
title "Descriptive Statistics for SBP versus Age";
data SBP;
input X Y;
  label X = 'Age'
        Y = 'Blood Pressure';
cards;
42 130
46 115
...
75 165
;
proc UNIVARIATE data=SBP;
Normal;
Plots/
Plotsize = 26;
Var Y;
run;
```

NORMAL helps to test if Blood Pressure (Y) is normally distributed

PLOTS provides three useful graphs: **Stem and Leaf, Box Plot, and Q-Q Plot.**

Option **HISTOGRAM** can be added to obtain the fourth graph.

Plotsize can be changed

Similar to that used with Q-Q Plot in Regression

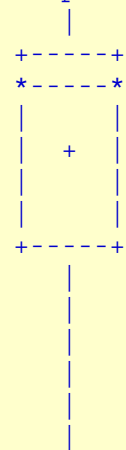
Variable=Y

Blood Pressure

```
Stem Leaf          #
 16 5              1
 16 022           3
 15 5668          4
 15 01            2
 14 8             1
 14
 13
 13 0             1
 12 5             1
 12
 11 5            1
 11
 10
 10 0            1
-----+-----+-----+-----+
```

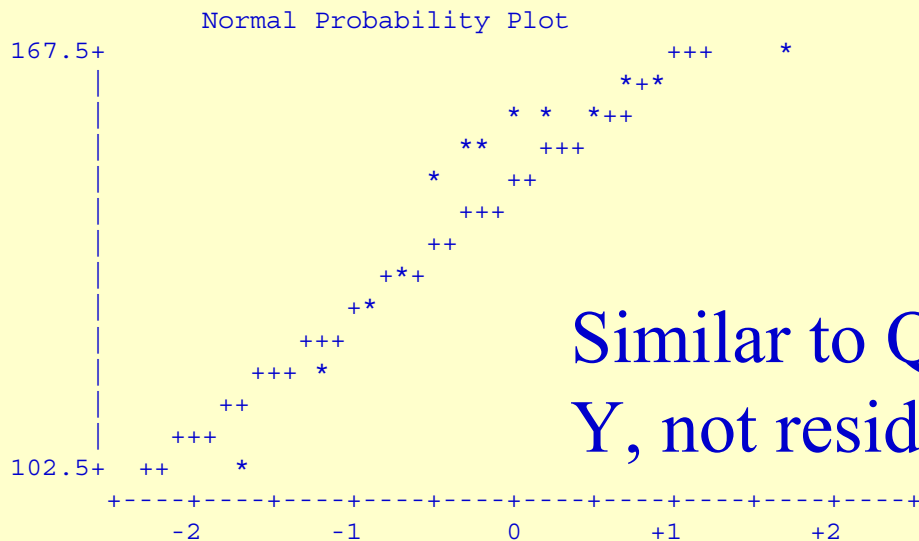
Multiply Stem.Leaf by 10**+1

Boxplot



There is a separate
PROC BOXPLOT
too!

RESULTING GRAPHS



Similar to Q-Q plot but plotting
Y, not residual on vertical axis

SET UP A TABLE OF DATA

```
options ls=79;  
title "Descriptive Statistics for SBP versus Age";  
data SBP;  
input X Y;  
    label X = 'Age'  
        Y = 'Blood Pressure';
```

```
cards;  
42 130  
46 115  
...  
75 165  
;
```

```
proc sql;
```

```
    create table temp as  
    select *, x - mean(x) as xdif, y - mean(y) as ydif, (x - mean(x))*( y - mean(y))  
        as crp, (x - mean(x))*(x - mean(x)) as sqdevx, (y - mean(y))*(y - mean(y))  
        as sqdevy from SBP;
```

```
proc print data = temp;
```

```
    var x y xdif ydif crp sqdevx sqdevy;
```

```
run;
```

This is an optional step.

Together **SQL** and **PRINT** would help to set up a Table just like **TABLE 1.1** on page 19 of the text book.

OBS	X	Y	XDIF	YDIF	CRP	SQDEVX	SQDEVY
1	42	130	-23.6	-16.2	382.32	556.96	262.44
2	46	115	-19.6	-31.2	611.52	384.16	973.44
3	42	148	-23.6	1.8	-42.48	556.96	3.24
4	71	100	5.4	-46.2	-249.48	29.16	2134.44
5	80	156	14.4	9.8	141.12	207.36	96.04
6	74	162	8.4	15.8	132.72	70.56	249.64
7	70	151	4.4	4.8	21.12	19.36	23.04
8	80	156	14.4	9.8	141.12	207.36	96.04
9	85	162	19.4	15.8	306.52	376.36	249.64
10	72	158	6.4	11.8	75.52	40.96	139.24
11	64	155	-1.6	8.8	-14.08	2.56	77.44
12	81	160	15.4	13.8	212.52	237.16	190.44
13	41	125	-24.6	-21.2	521.52	605.16	449.44
14	61	150	-4.6	3.8	-17.48	21.16	14.44
15	75	165	9.4	18.8	176.72	88.36	353.44

TABLE 1.1 (Page 19, Text)

SQL (Structured Query Language) is a popular standardized language that retrieves and updates data in tables and views based on those tables.

From TA's HANDOUT #2

data example;

```
input ID gender $ exam1 exam2 hwgrade $ ;
```

```
final=(exam1+exam2)/2;
```

```
If final GE 0 and final LT 60 then grade='F';
```

```
  Else if final GE 60 and final LT 75 then grade='C';
```

```
  Else if final GE 75 and final LT 90 then grade='B';
```

```
  Else if final GE 90 then grade='A';
```

```
datalines;
```

```
30004 M 90 96 A
```

```
30001 F 92 88 A
```

```
30002 M 78 88 B
```

```
30005 M 88 92 B
```

```
30006 F 87 89 B
```

```
30003 M 67 78 C;
```

```
run;
```

```
Proc sort data=example out=sort;
```

```
  by ID;
```

```
run;
```

```
Proc sort data=example out=sort1;
```

```
  by grade;
```

```
run;
```

```
proc print data=sort1 noobs;
```

```
  Title 'Roster ordered in student ID';
```

```
  Var ID exam1 exam2 final hwgrade grade;
```

```
run;
```

Data as non-numerical

Define/Calculate New Variable

Categorize continuous measurement

Important Part:

HOW to sort data

CORRELATION (& Scatter Diagram)

```
options ls=79;  
title "Descriptive Statistics for SBP versus Age";  
data SBP;  
input X Y;  
  label X = 'Age'  
        Y = 'Blood Pressure';  
cards;  
42 130  
46 115  
...  
75 165  
;
```

```
proc CORR data=SBP;
```

```
run;
```

```
proc plot data=SBP;
```

```
  plot y*x='*';
```

```
run;
```

Proc CORR gives the coefficient of correlation r (& the p -value)

Proc PLOT provides the Scatter Diagram; could choose symbol to plot.

Specify Notation for the graph

Simple Statistics

Variable	N	Mean	Std Dev	Sum
X	15	65.600000	15.592123	984.000000
Y	15	146.200000	19.479660	2193.000000

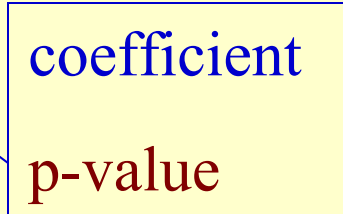
Simple Statistics

Variable	Minimum	Maximum	Label
X	41.000000	85.000000	Age
Y	100.000000	165.000000	Blood Pressure

OUTPUT

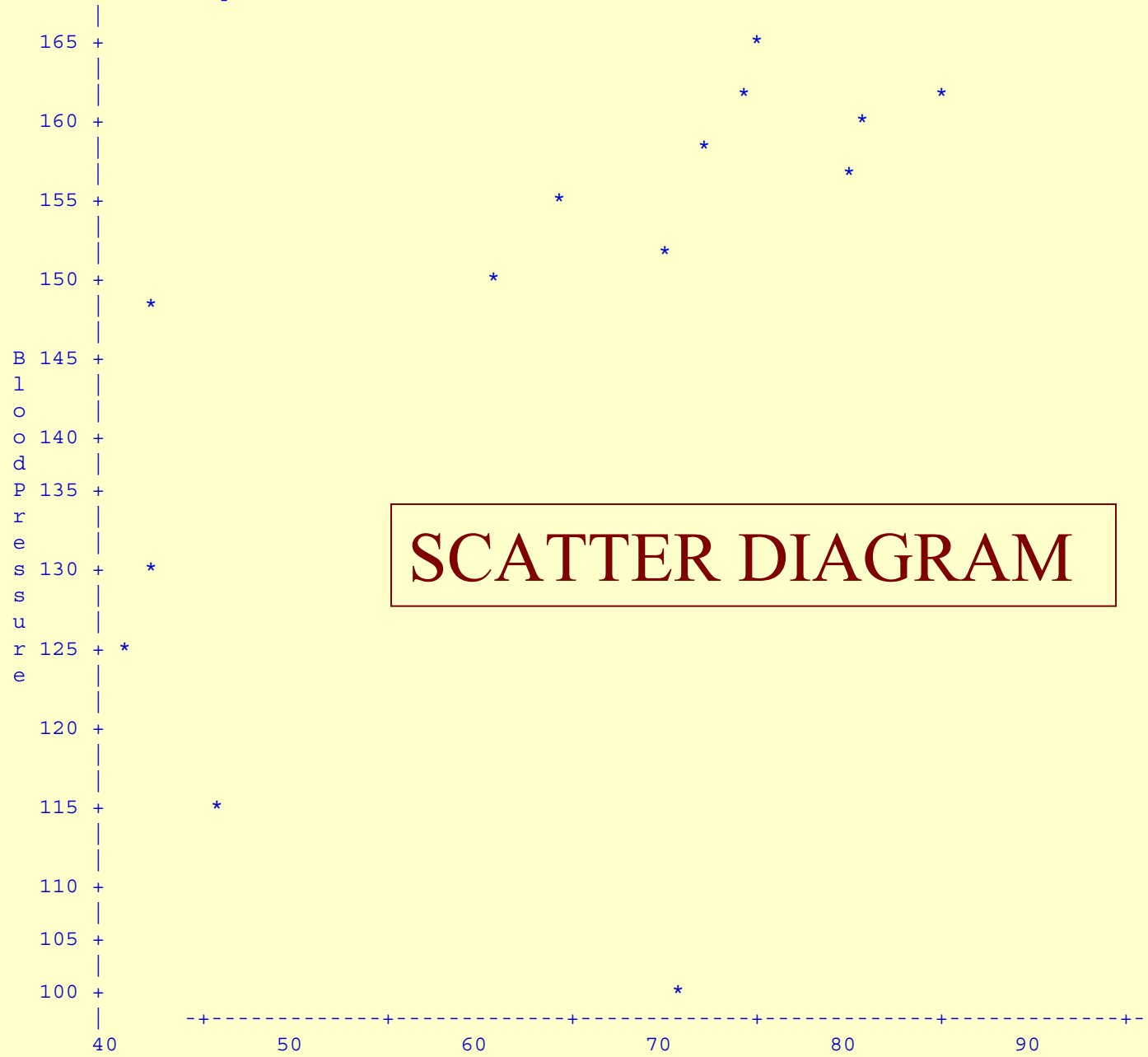
Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 15

	X	Y
X	1.00000	0.56422
Age	0.0	0.0285
Y	0.56422	1.00000
Blood Pressure	0.0285	0.0



Note: results are symmetric

Plot of Y*X. Symbol used is '*'.



SCATTER DIAGRAM

SIMPLE LINEAR REGRESSION (& Scatter Diagram)

```
options ls=79;  
title "Descriptive Statistics for SBP versus Age";  
data SBP;  
input X Y;  
  label X = 'Age'  
        Y = 'Blood Pressure';  
cards;  
42 130  
46 115  
...  
75 165  
;
```

Proc REG is the most basic one; will add in more options

PLOT provides the Scatter Diagram; could choose symbol to plot.

```
proc REG data = SBP;
```

```
model y = x; ←————— Key: Model Statement
```

```
plot y*x='+';
```

CORR and REG provide the same Scatter Diagram (“plot” option)

```
run;
```

PARAMETER ESTIMATES

Testing for Zero Intercept
(usually not needed)

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T
INTERCEP	1	99.958515	19.25516927	5.191	0.0002
X (Age)	1	0.704901	0.28607866	2.464	0.0285

Variable	DF	Variable Label
INTERCEP	1	Intercept
X	1	Age

Slope

Testing for Zero Slope
(i.e. Independence)

ANALYSIS OF VARIANCE

Testing for Independence

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	1691.19774	1691.19774	6.071	0.0285
Error	13	3621.20226	278.55402		
Total	14	5312.40000			
Root MSE		16.68994			
Dep Mean		146.20000			
C.V.		11.41583			
			R-square	0.3183	

From R^2 & slope, obtain "r"

MSE & its square root

SET UP ANOTHER TABLE

```
options ls=79;
title "Descriptive Statistics for SBP versus Age";
data SBP;
input X Y;
  label X = 'Age'
        Y = 'Blood Pressure';
cards;
42 130
...
75 165
;
```

This program, a few extra steps and **proc PRINT** would help to set up a Table just like **TABLE 1.2 on page 22** of the text book (note the use of **NOPRINT**).

```
proc reg data = SBP noprint;
```

```
model y = x;
```

```
output out=SBP1 p=yhat r=residual;
```

```
run;
```

```
data SBP2;
```

```
set SBP1;
```

```
rsq=residual**2;
```

```
run;
```

```
proc print data=SBP2;
```

```
var x y yhat residual rsq;
```

```
run;
```

Setup a file to keep the results

Residual

Define Squared Residuals

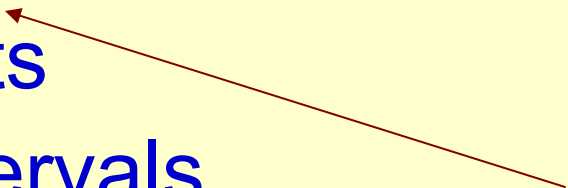
Fitted Value

OBS	X	Y	YHAT	RESIDUAL	RSQ
1	42	130	129.564	0.4357	0.19
2	46	115	132.384	-17.3839	302.20
3	42	148	129.564	18.4357	339.87
4	71	100	150.006	-50.0065	2500.65
5	80	156	156.351	-0.3506	0.12
6	74	162	152.121	9.8788	97.59
7	70	151	149.302	1.6984	2.88
8	80	156	156.351	-0.3506	0.12
9	85	162	159.875	2.1249	4.52
10	72	158	150.711	7.2886	53.12
11	64	155	145.072	9.9278	98.56
12	81	160	157.055	2.9445	8.67
13	41	125	128.859	-3.8594	14.90
14	61	150	142.957	7.0425	49.60
15	75	165	152.826	12.1739	148.20

(TABLE 1.2, page 22)

USEFUL OPTIONS FROM PROC REG

- **R**: Analysis of residuals
- **P**: computing predicted values (i.e. fitted)
- **COVB**: Var-Cov matrix of regression coefficients
- **CLM**: Confidence Intervals of mean responses
- **CLI**: Conf Intervals of new individual responses


$$\begin{bmatrix} s^2(b_0) & s(b_0, b_1) \\ s(b_0, b_1) & s^2(b_1) \end{bmatrix}$$

ANALYSIS OF RESIDUALS

```
options ls=79;
title "Descriptive Statistics for SBP versus Age";
data SBP;
input X Y;
    label X = 'Age'
        Y = 'Blood Pressure';
cards;
42 130
...
75 165
;
proc reg data = SBP noprint;
model y = x/R;
run;
```

This short program achieves the same thing and more; it helps to set up a Table just like **TABLE 1.2 on page 22** of the text book and student residuals – plus all regression analysis results.

Standard Results come first

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	1691.19774	1691.19774	6.071	0.0285
Error	13	3621.20226	278.55402		
C Total	14	5312.40000			

Root MSE 16.68994 R-square 0.3183
Dep Mean 146.20000 Adj R-sq 0.2659
C.V. 11.41583

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T
INTERCEP	1	99.958515	19.25516927	5.191	0.0002
X	1	0.704901	0.28607866	2.464	0.0285

... then Results for Residuals

Obs	Dep Var Y	Predict Value	Std Err Predict	Std Err Residual	Student Residual	
1	130.0	129.6	8.010	0.4357	14.642	0.030
2	115.0	132.4	7.072	-17.3839	15.118	-1.150
3	148.0	129.6	8.010	18.4357	14.642	1.259
4	100.0	150.0	4.578	-50.0065	16.050	-3.116
5	156.0	156.4	5.962	-0.3506	15.589	-0.022
6	162.0	152.1	4.934	9.8788	15.944	0.620
7	151.0	149.3	4.489	1.6984	16.075	0.106
8	156.0	156.4	5.962	-0.3506	15.589	-0.022
9	162.0	159.9	7.027	2.1249	15.139	0.140
10	158.0	150.7	4.682	7.2886	16.020	0.455
11	155.0	145.1	4.334	9.9278	16.118	0.616
12	160.0	157.1	6.163	2.9445	15.510	0.190
13	125.0	128.9	8.252	-3.8594	14.507	-0.266
14	150.0	143.0	4.506	7.0425	16.070	0.438
15	165.0	152.8	5.080	12.1739	15.898	0.766

These are Studentized Residuals



EXAMPLE: Option COVB

```
options ls=79;
title "Descriptive Statistics for SBP
versus Age";
data SBP;
input X Y;
    label X = 'Age'
           Y = 'Blood Pressure';
cards;
42 130
46 115
...
75 165
;
proc REG data = SBP;
model y = x/COVB;
run;
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T
INTERCEP	1	99.958515	19.25516927	5.191	0.0002
X	1	0.704901	0.28607866	2.464	0.0285

Covariance of Estimates

COVB	INTERCEP	X
INTERCEP	370.76154379	-5.368769448
X	-5.368769448	0.0818409977

$\text{Var}(b_0)$

$\text{Var}(b_1)$

$\text{Cov}(b_0, b_1)$

Intercept
Age

EXAMPLE: Option CLM

```
options ls=79;
title "Descriptive Statistics for SBP
versus Age";
data SBP;
input X Y;
    label X = 'Age'
           Y = 'Blood Pressure';
cards;
42 130
46 115
...
75 165
;
proc REG data = SBP;
model y = x/CLM;
run;
```

Obs	Dep Var Y	Predict Value	Std Err Predict	Lower95% Mean	Upper95% Mean	Residual
1	130.0	129.6	8.010	112.3	146.9	0.4357
2	115.0	132.4	7.072	117.1	147.7	-17.3839
3	148.0	129.6	8.010	112.3	146.9	18.4357
4	100.0	150.0	4.578	140.1	159.9	-50.0065
5	156.0	156.4	5.962	143.5	169.2	-0.3506
6	162.0	152.1	4.934	141.5	162.8	9.8788
7	151.0	149.3	4.489	139.6	159.0	1.6984
8	156.0	156.4	5.962	143.5	169.2	-0.3506
9	162.0	159.9	7.027	144.7	175.1	2.1249
10	158.0	150.7	4.682	140.6	160.8	7.2886
11	155.0	145.1	4.334	135.7	154.4	9.9278
12	160.0	157.1	6.163	143.7	170.4	2.9445
13	125.0	128.9	8.252	111.0	146.7	-3.8594
14	150.0	143.0	4.506	133.2	152.7	7.0425
15	165.0	152.8	5.080	141.9	163.8	12.1739

Sum of Residuals 0
Sum of Squared Residuals 3621.2023

EXAMPLE: Option CLI

```
options ls=79;
title "Descriptive Statistics for SBP
versus Age";
data SBP;
input X Y;
    label X = 'Age'
           Y = 'Blood Pressure';
cards;
42 130
46 115
...
75 165
;
proc REG data = SBP;
model y = x/CLI;
run;
```

Output Statistics (from CLI)

Versus: [112.3,146.9]
Under CLM

Obs	Dep Var Y	Predicted Value	Std Error Mean Predict	95% CL Predict	Residual
1	130.0000	129.5643	8.0095	89.5709 169.5578	0.4357
2	115.0000	132.3839	7.0718	93.2244 171.5435	-17.3839
3	148.0000	129.5643	8.0095	89.5709 169.5578	18.4357
4	100.0000	150.0065	4.5779	112.6183 187.3946	-50.0065
5	156.0000	156.3506	5.9616	118.0630 194.6382	-0.3506
6	162.0000	152.1212	4.9341	114.5221 189.7202	9.8788
7	151.0000	149.3016	4.4894	111.9635 186.6396	1.6984
8	156.0000	156.3506	5.9616	118.0630 194.6382	-0.3506
9	162.0000	159.8751	7.0265	120.7535 198.9966	2.1249
10	158.0000	150.7114	4.6821	113.2630 188.1598	7.2886
11	155.0000	145.0722	4.3336	107.8201 182.3242	9.9278
12	160.0000	157.0555	6.1628	118.6195 195.4914	2.9445
13	125.0000	128.8594	8.2521	88.6365 169.0824	-3.8594
14	150.0000	142.9575	4.5058	105.6102 180.3047	7.0425
15	165.0000	152.8261	5.0795	115.1367 190.5154	12.1739

Sum of Residuals

0

Sum of Squared Residuals

3621.20226

Note: wider Intervals

```
proc reg data=example;  
model y=x/alpha=0.01 cli clm;  
run;
```

New Important Part:

Set 99% CI instead of 95%

Q-Q PLOT

```
proc REG data = SBP;  
model y = x/CLI;  
Plot r.*nqq./  
      noline mse cframe = ligr;  
Run;
```

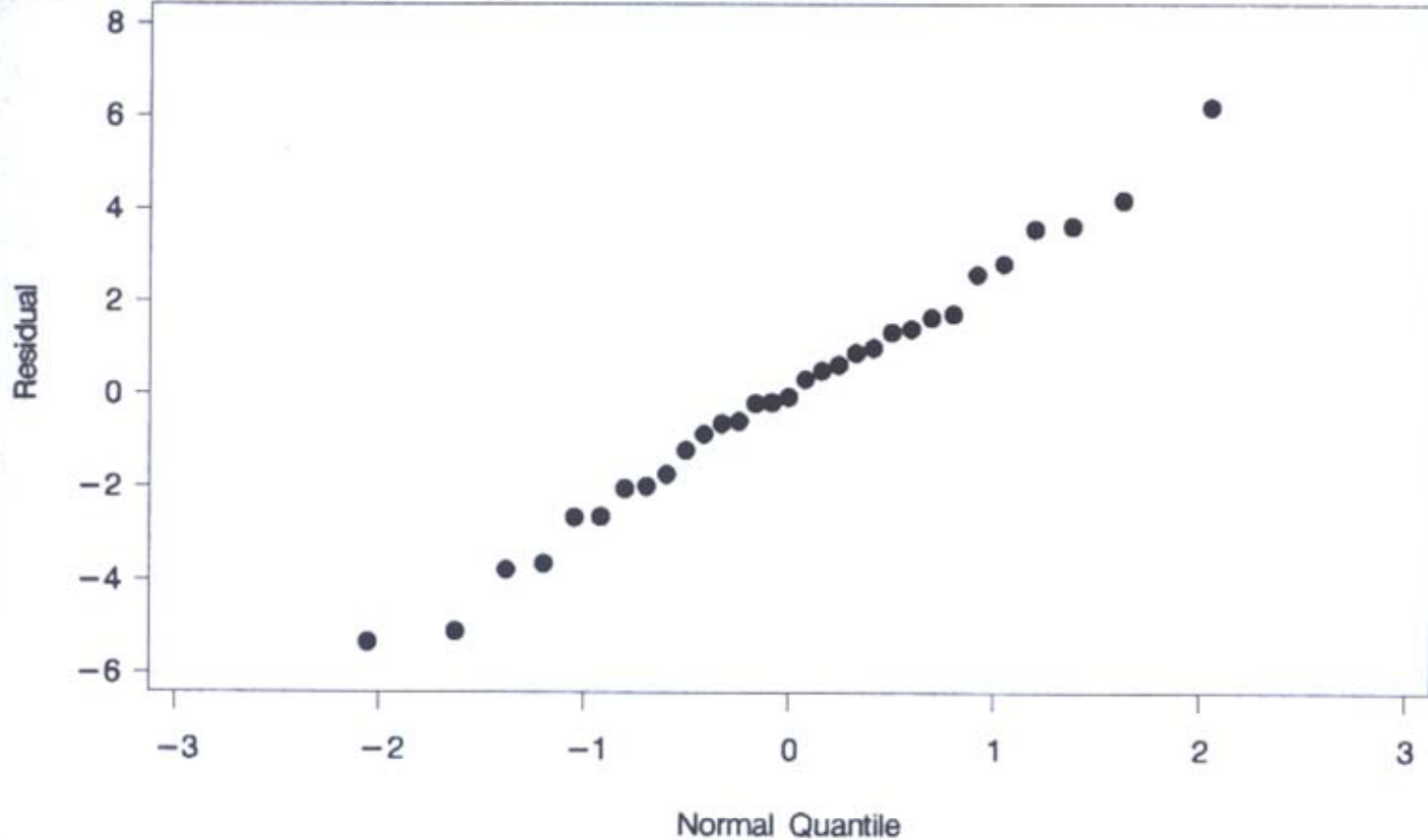
See pages 3020-3021

SAS Manual, vol. 3

Output 55.8.2. Normal Quantile-Quantile Plot for the Residuals

QQ Plot

'Best' Two-Parameter Model: Oxygen = 82.422 - 3.3106 RunTime



N
31
Rsq
0.7434
AdjRsq
0.7345
RMSE
2.7448
MSE
7.5338

Note: Correlation “r” needed for the “test” can be obtained

MORE PLOTS

```
options ls=79;
title "Descriptive Statistics for SBP versus
Age";
data SBP;
input X Y;
    label X = 'Age'
        Y = 'Blood Pressure';
cards;
42 130
46 115
...
75 165
;
proc reg data = SBP;
model Y = X;
plot r.*X="+" r.*p.="*";
    student.*X="+" student.*p.="*";
run;
```

“r.” & “p.” are default notations

2 graphs are requested:

(1) Residual versus X

(2) Residual versus Fitted Value

Actually, Four graphs:
2 for residuals and 2 for
studentized residuals

MORE PLOTS

```
options ls=79;
title "Descriptive Statistics for SBP versus
Age";
data SBP;
input X Y;
    label X = 'Age'
        Y = 'Blood Pressure';
cards;
42 130
46 115
...
75 165
;
proc reg data = SBP;
model Y = X;
plot r.*X="+" r.*p.="*"
    student.*X="+" student.*p.="*"/
hplots=2;
run;
```

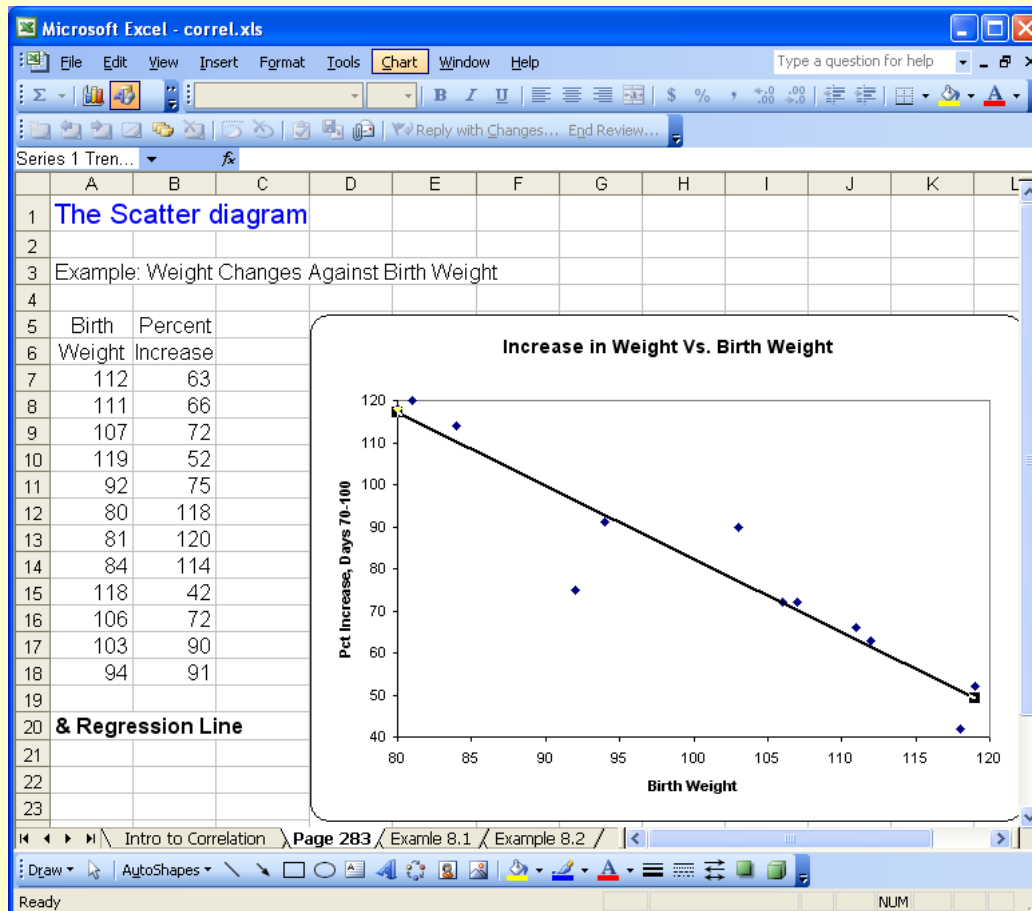
Same 4 graphs are requested:

- (1) Residual versus X
- (2) Residual versus Fitted Value
- (3) Student Residual versus X
- (4) Student residual versus Fitted Value

Two graphs are on the same page

Excel: REGRESSION LINE

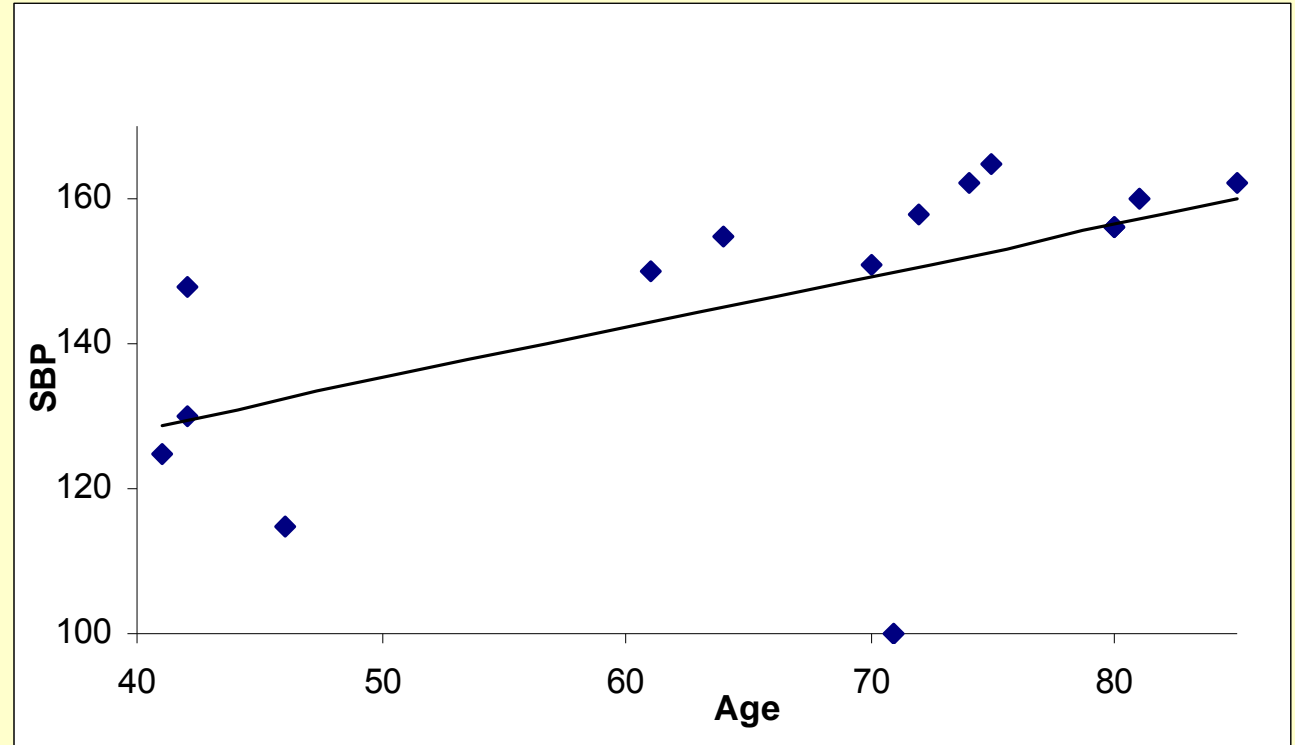
Steps: (a) Click on the new Chart (scatter diagram) to make it active, (b) Click on *Chart* (on the top row menu), (c) a box appears to let you choose “Add Trendline”



Example #2: Age and SBP

$n = 15$, $X = \text{AGE}$ (Years), $Y = \text{Systolic Blood Pressure}$ (in mm of Hg)

Age (x)	SBP (y)
42	130
46	115
42	148
71	100
80	156
74	162
70	151
80	156
85	162
72	158
64	155
81	160
41	125
61	150
75	165

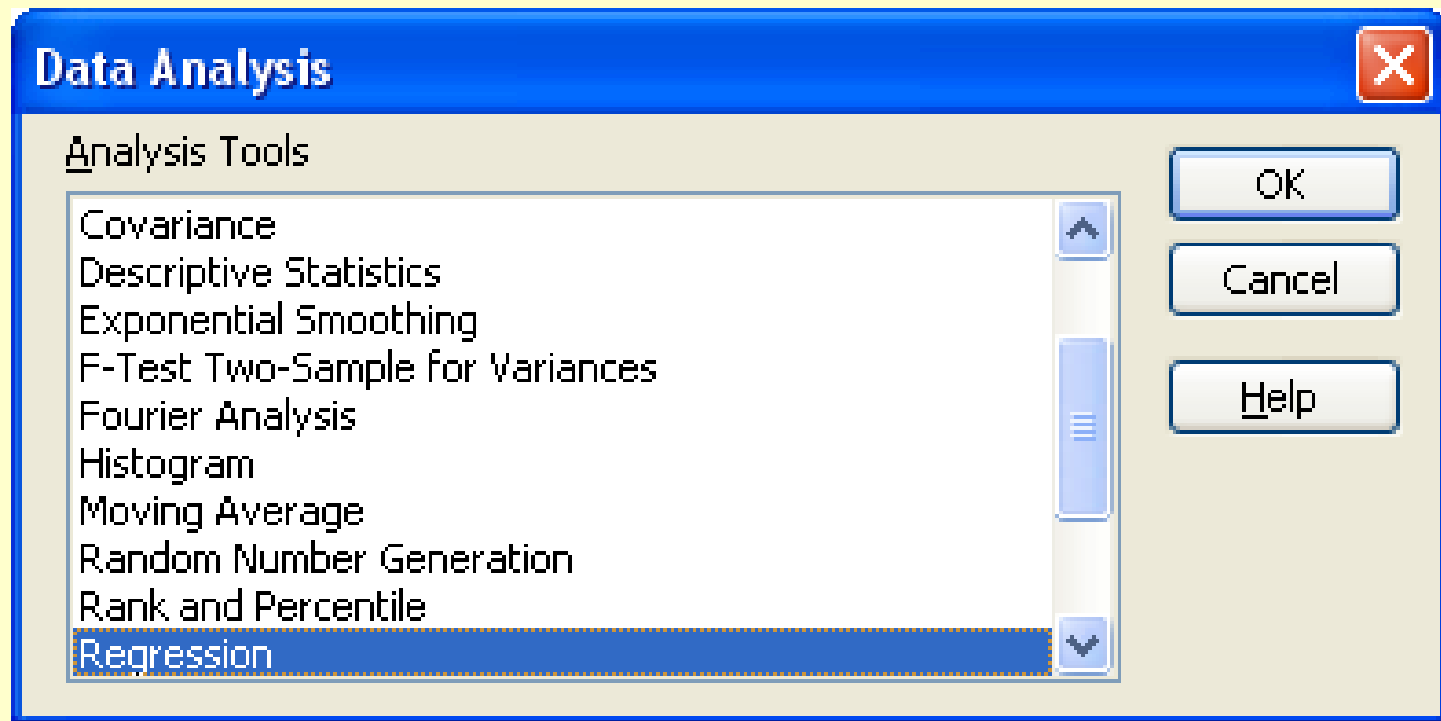


SBP versus AGE

(AGE is Predictor, SBP is Response)

Excel: ANALYSIS

(1) click the *Tools* then (2) *Data Analysis*; among functions available, choose *Regression*.



A box appears, use the cursor to fill in the ranges of Y and X's. The results include all items needed, including regression estimates of coefficients, their standard errors, and their 95% confidence intervals. And much more

Regression

Input

Input Y Range:

Input X Range:

Labels Constant is Zero

Confidence Level: %

Output options

Output Range:

New Worksheet Ply:

New Workbook

Residuals

Residuals Residual Plots

Standardized Residuals Line Fit Plots


Normal Probability


Normal Probability Plots

OK
Cancel
Help

Regression ✕

Input


Input Y Range: 

Input X Range: 

Labels Constant is Zero

Confidence Level: %

Output options

Output Range: 

New Worksheet Ply:

New Workbook

Residuals

Residuals Residual Plots

Standardized Residuals Line Fit Plots

Normal Probability

Normal Probability Plots

You can check as many boxes as you want;
can change the degree of confidence too.

ANOVA & SUMMARY

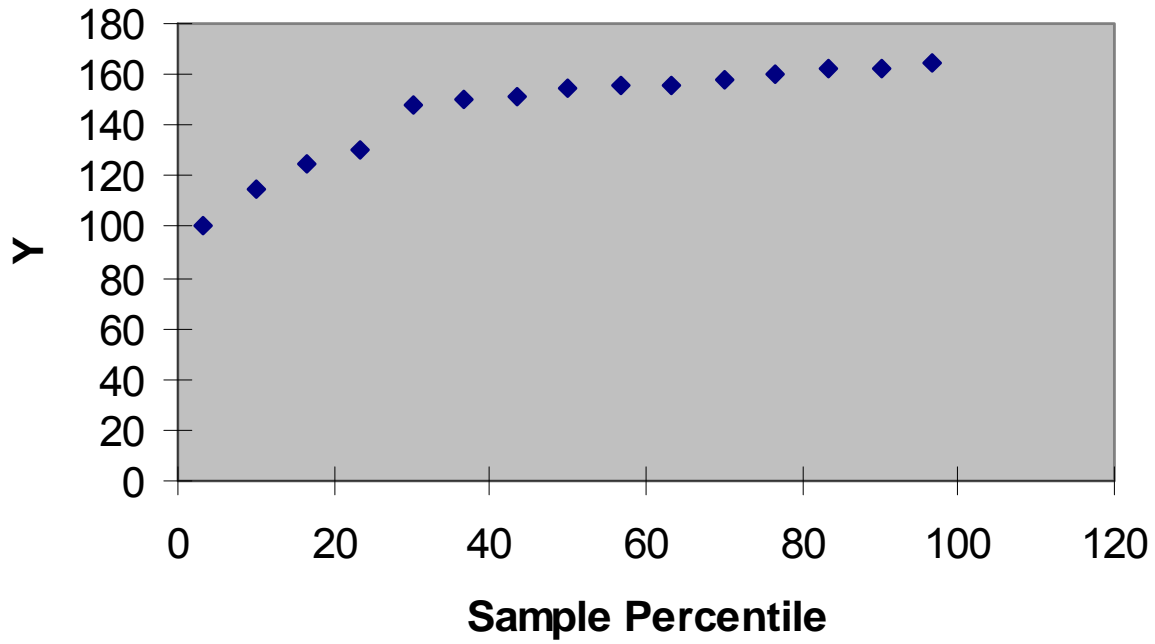
SUMMARY OUTPUT	
<i>Regression Statistics</i>	
Multiple R	0.564224332
R Square	0.318349097
Adjusted R Square	0.265914412
Standard Error	16.68993768
Observations	15

ANOVA	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	1691.198	1691.197744	6.071346	0.028453563
Residual	13	3621.202	278.5540197		
Total	14	5312.4			

PARAMETER ESTIMATES & CI

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	99.95851451	19.25516927	5.1912561	0.00017371	58.36025042	141.5567786
X Variable 1	0.704900693	0.286078656	2.4640101	0.02845356	0.086865332	1.322936055

Normal Probability Plot



Readings & Exercises

- **Readings: none**
- **Exercises: Try to reproduced all or most of the results you have seen from this review section using the “birth weight” data set (at right)**
- **Due as Homework: none.**

x (oz)	y (%)
112	63
111	66
107	72
119	52
92	75
80	118
81	120
84	114
118	42
106	72
103	90
94	91

Due As Homework

11.1 Find the number “a” which minimizes the following sum.
What is “a” called?

$$Q = \sum_{i=1}^n (x_i - a)^2$$

11.2 We obtain “least squares estimates” of slope & intercept by solving the following system of two linear equations:

$$Q = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\delta Q}{\delta \beta_0} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\delta Q}{\delta \beta_1} = -2 \sum_{i=1}^n x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

How do we know if the solutions minimize the sum of squares Q, rather than maximizing it ?