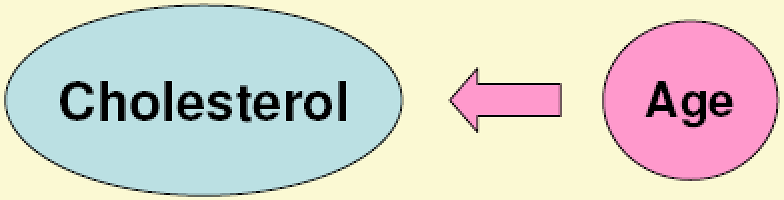
Simple Linear Regression

Azmi Mohd Tamil

Simple Linear Regression

- To determine the relationship between age and blood cholesterol level



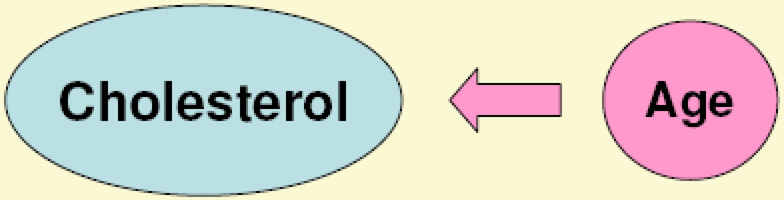
► Here, we may use either '<u>correlation analysis</u>' or '<u>regression analysis</u>', as both cholesterol and age are numerical variables.

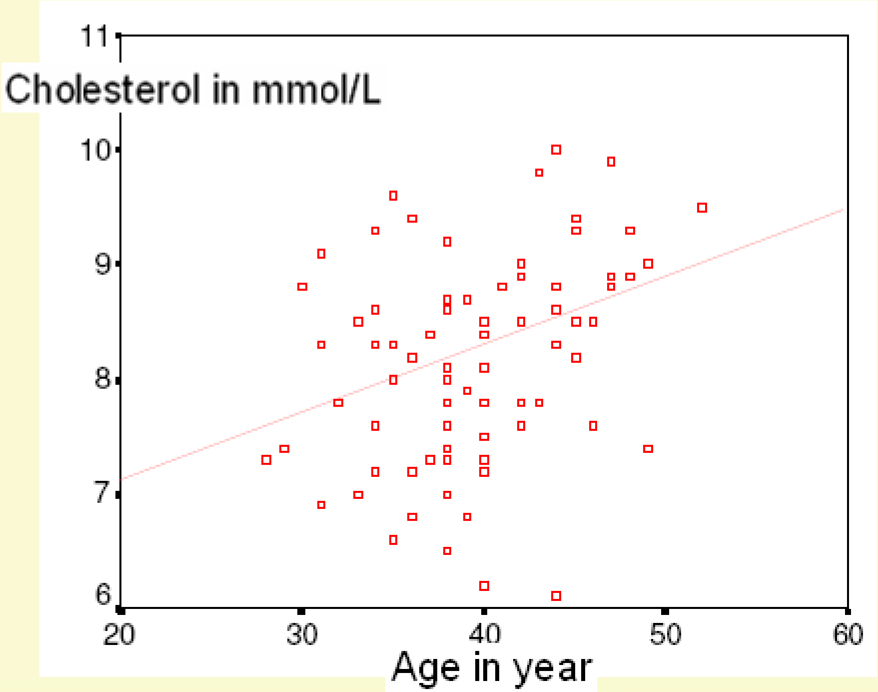
► *Correlation* can give the strength of relationship, but *regression* can describe the relationship in more detail.

► In above example, if we decide to do <u>*regression*</u>, cholesterol will be our outcome (dependent) variable, because age may determine cholesterol but cholesterol cannot determine age.

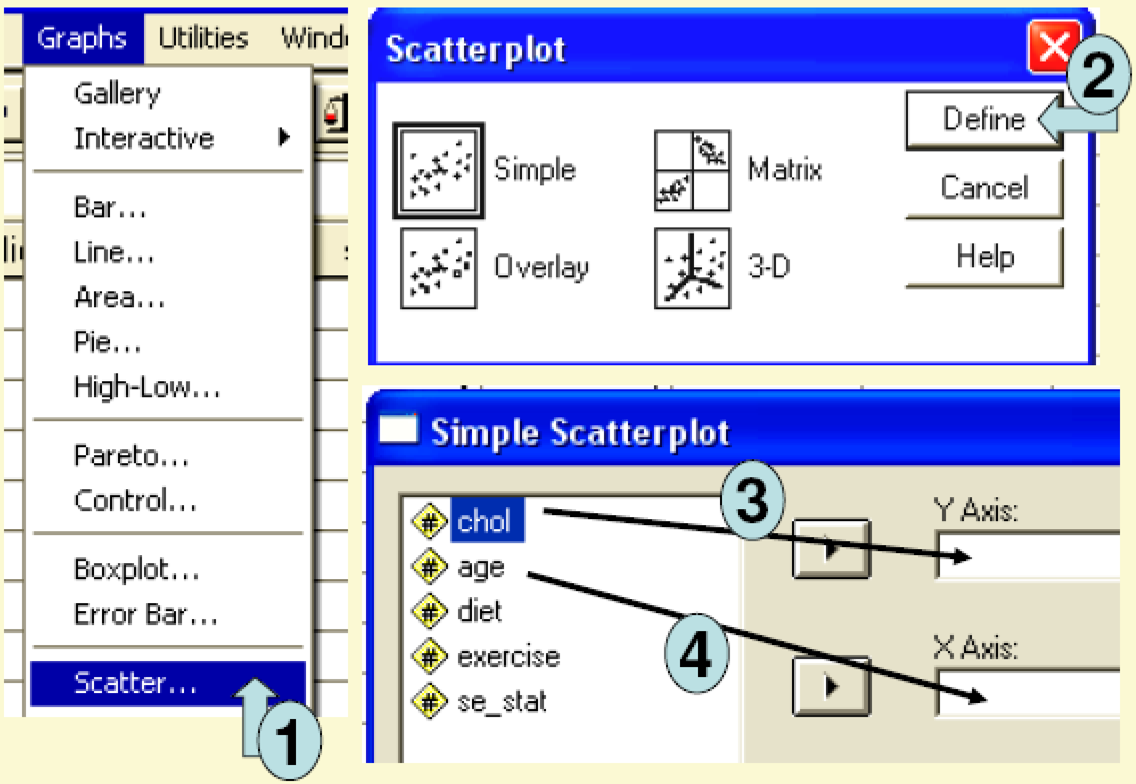
Simple Linear Regression

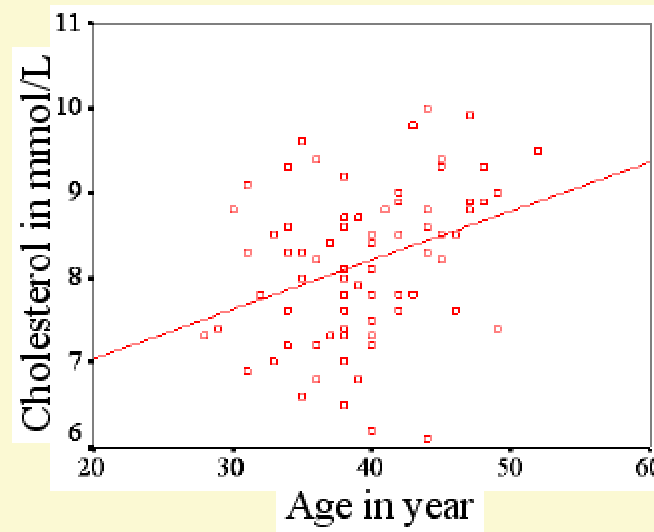
- To determine the relationship between age and blood cholesterol level



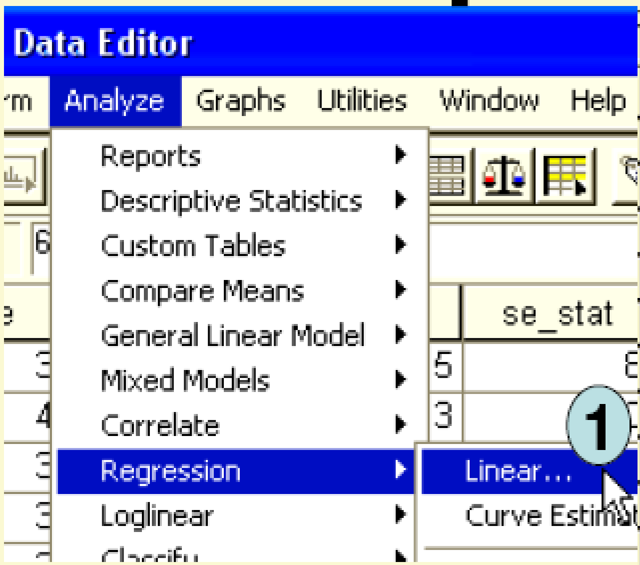


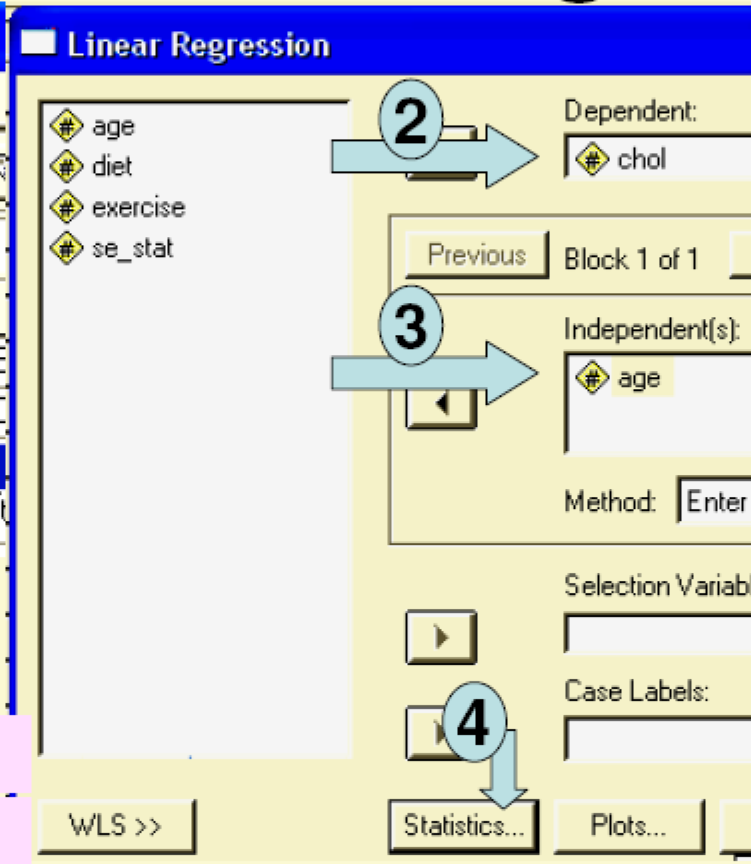
Simple Linear Regression

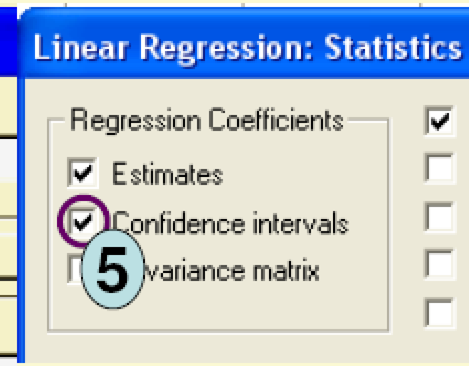




Simple Linear Regression







Y = a + bX

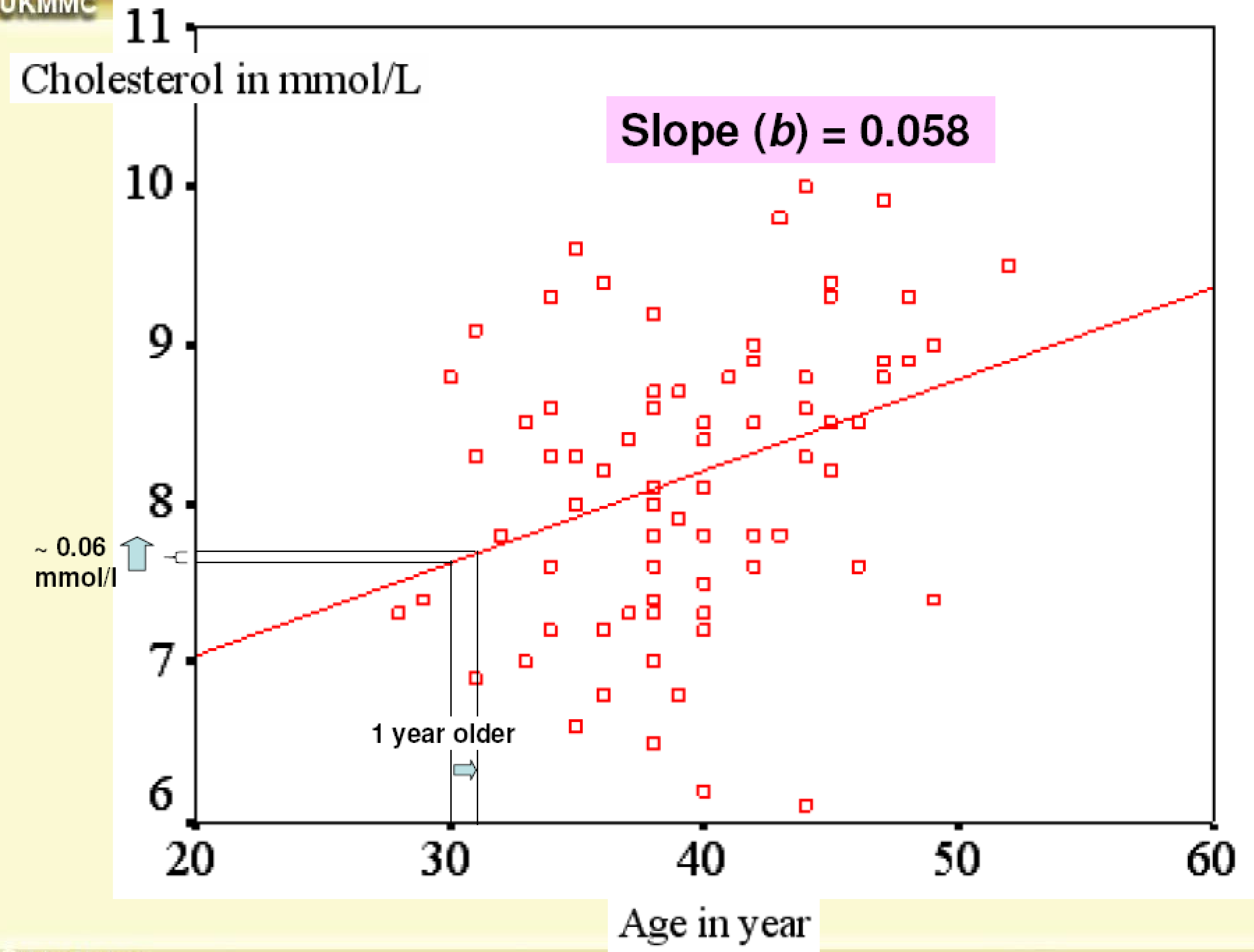
Chol = 5.9 + (0.058*age)

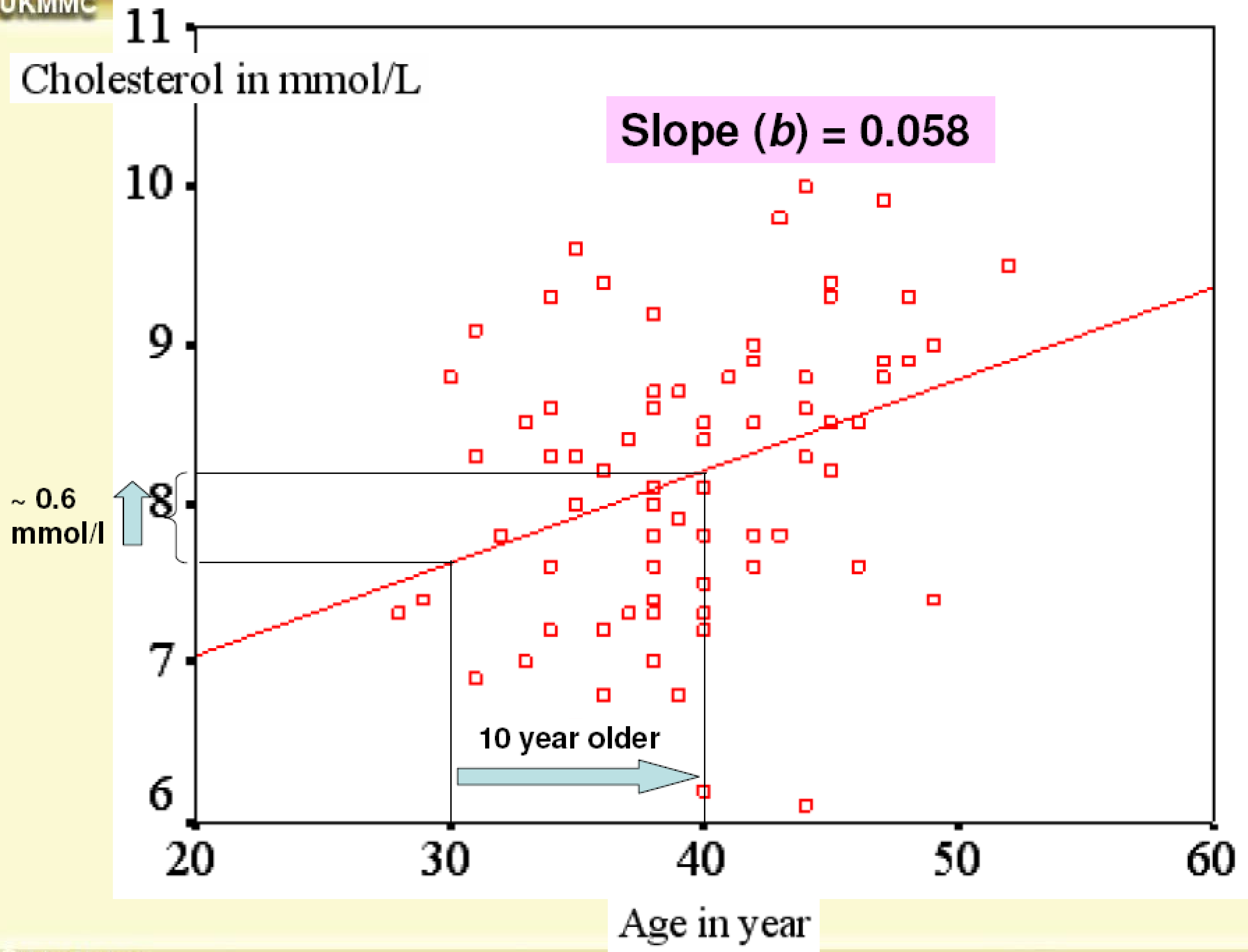
Coefficients[a]

$H_o: \beta=0$

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	*P* value Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B Upper Bound
1	(Constant)	5.895	.735		8.026	.000	4.434	7.357
	AGE age in year	5.776E-02	.018	.331	3.134	.002	.021	.094

a. Dependent Variable: CHOL cholesterol in mmol/L


Cholesterol in mmol/L
Slope (b) = 0.058
~ 0.06
mmol/l
1 year older
Age in year
11
10
9
8
7
6
20
30
40
50
60

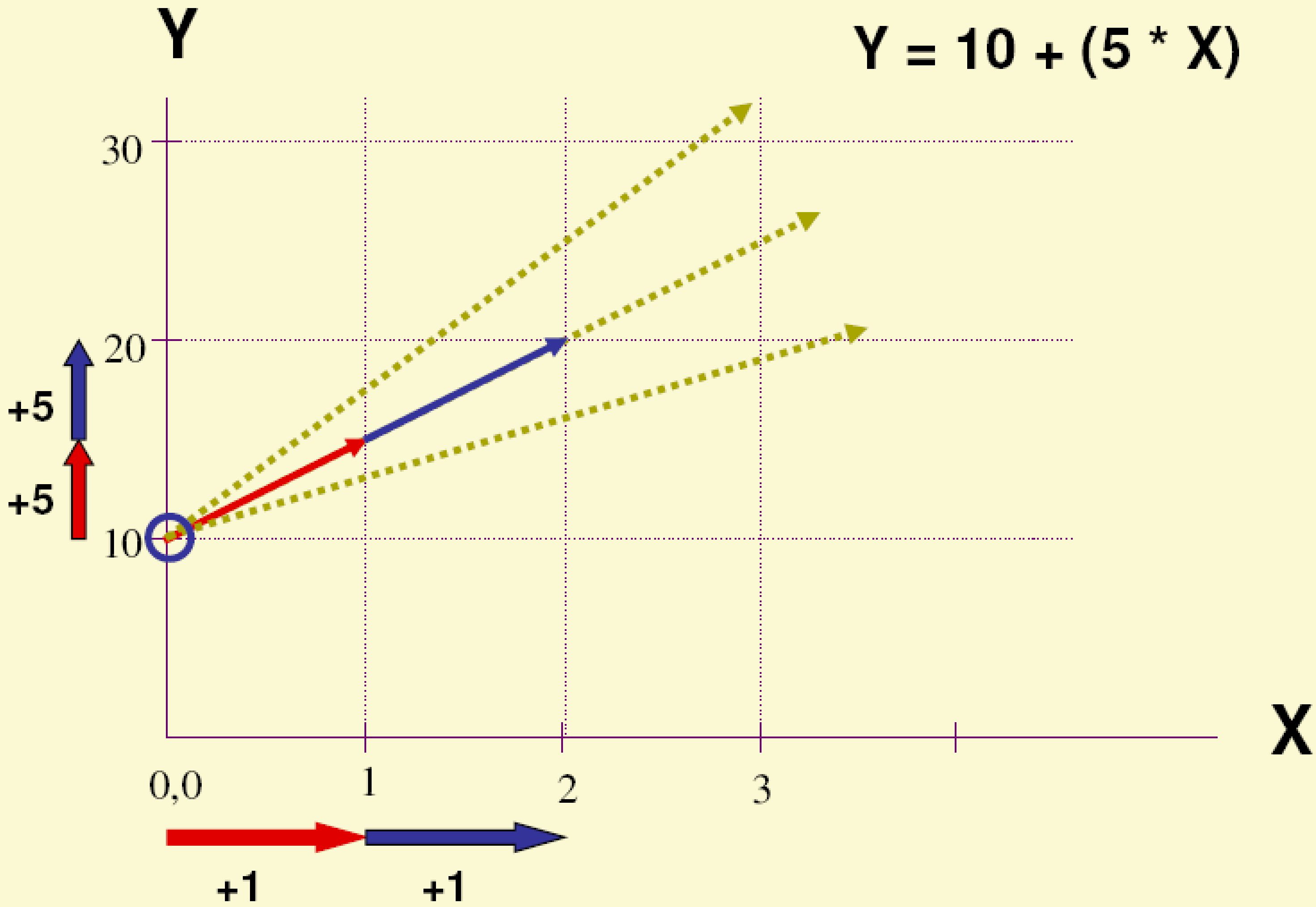

Cholesterol in mmol/L
11
10
9
8
7
6
Slope (b) = 0.058
~ 0.6
mmol/l
10 year older
20
30
40
50
60
Age in year

The Linear line is described by the "Linear Equation".

$$Y = a + (b * X)$$

$$Y = \text{Constant} + (\text{slope} * X)$$

$$Y = 10 + (5 * X)$$



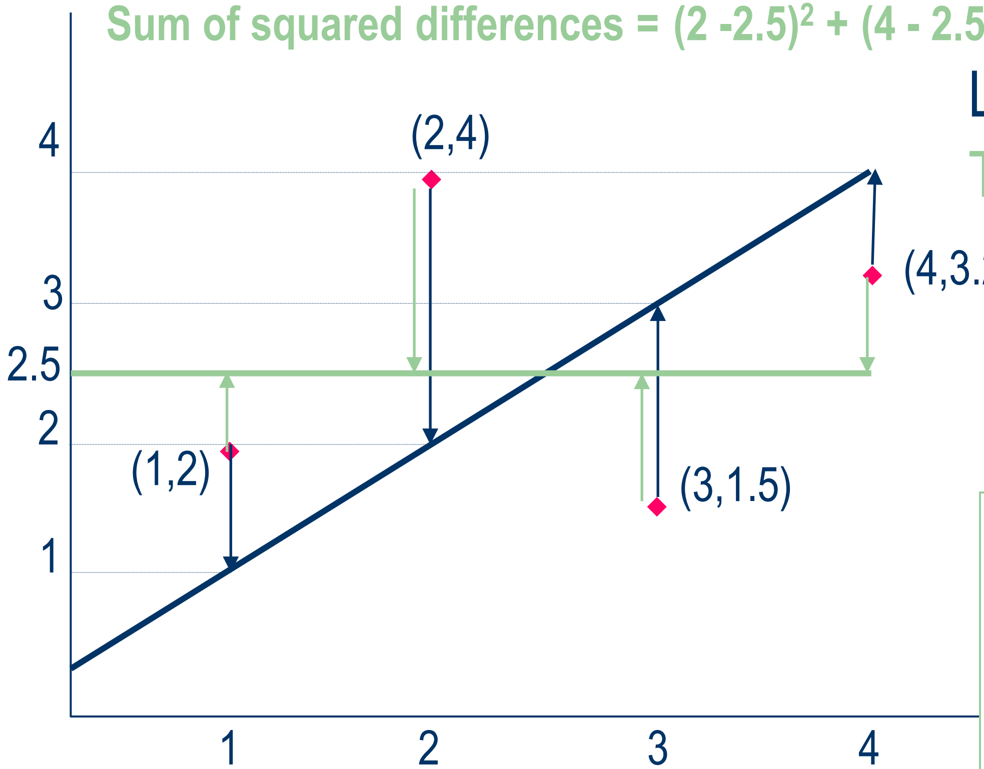
The Least Squares (Regression) Line

A good line is one that minimizes
the sum of squared differences between the points and the line.

The Least Squares (Regression) Line

Sum of squared differences = $(2 - 1)^2 + (4 - 2)^2 + (1.5 - 3)^2 + (3.2 - 4)^2 = 6.89$

Sum of squared differences = $(2 - 2.5)^2 + (4 - 2.5)^2 + (1.5 - 2.5)^2 + (3.2 - 2.5)^2 = 3.99$



Let us compare two lines

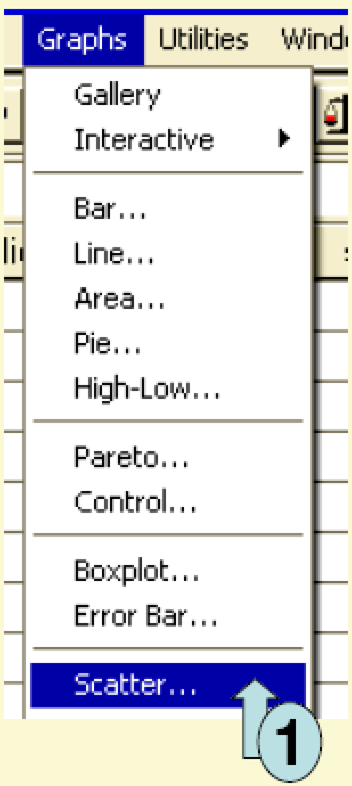
The second line is horizontal

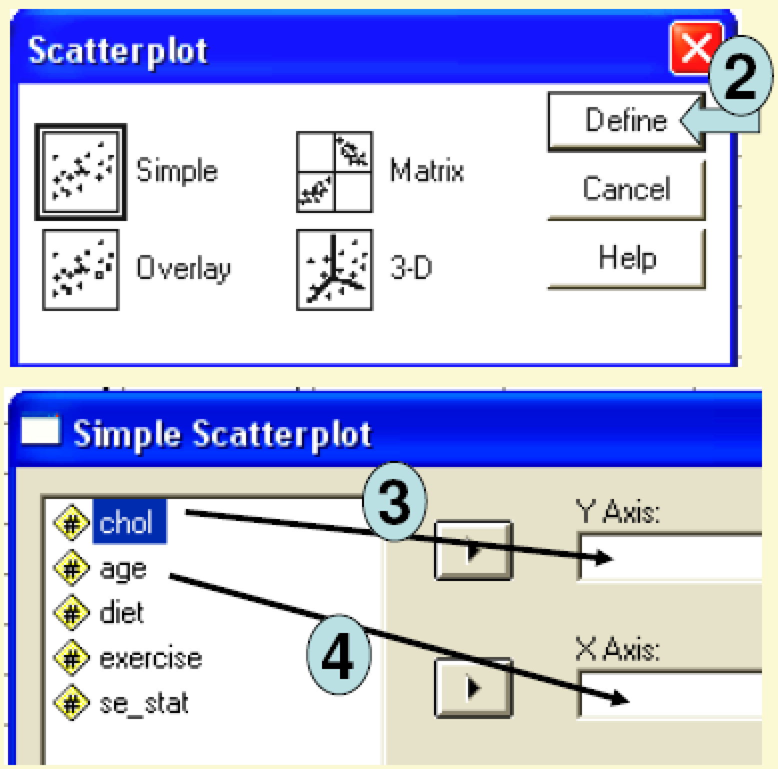
The smaller the sum of squared differences the better the fit of the line to the data.

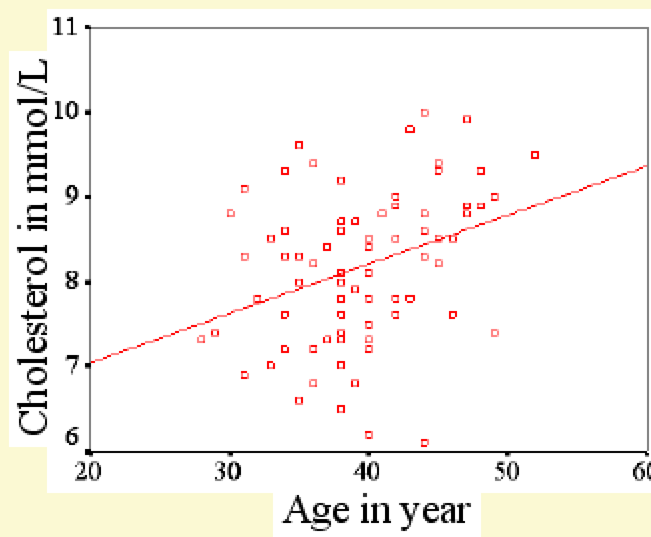
Step 2: Simple Linear Regression

Two main reasons:

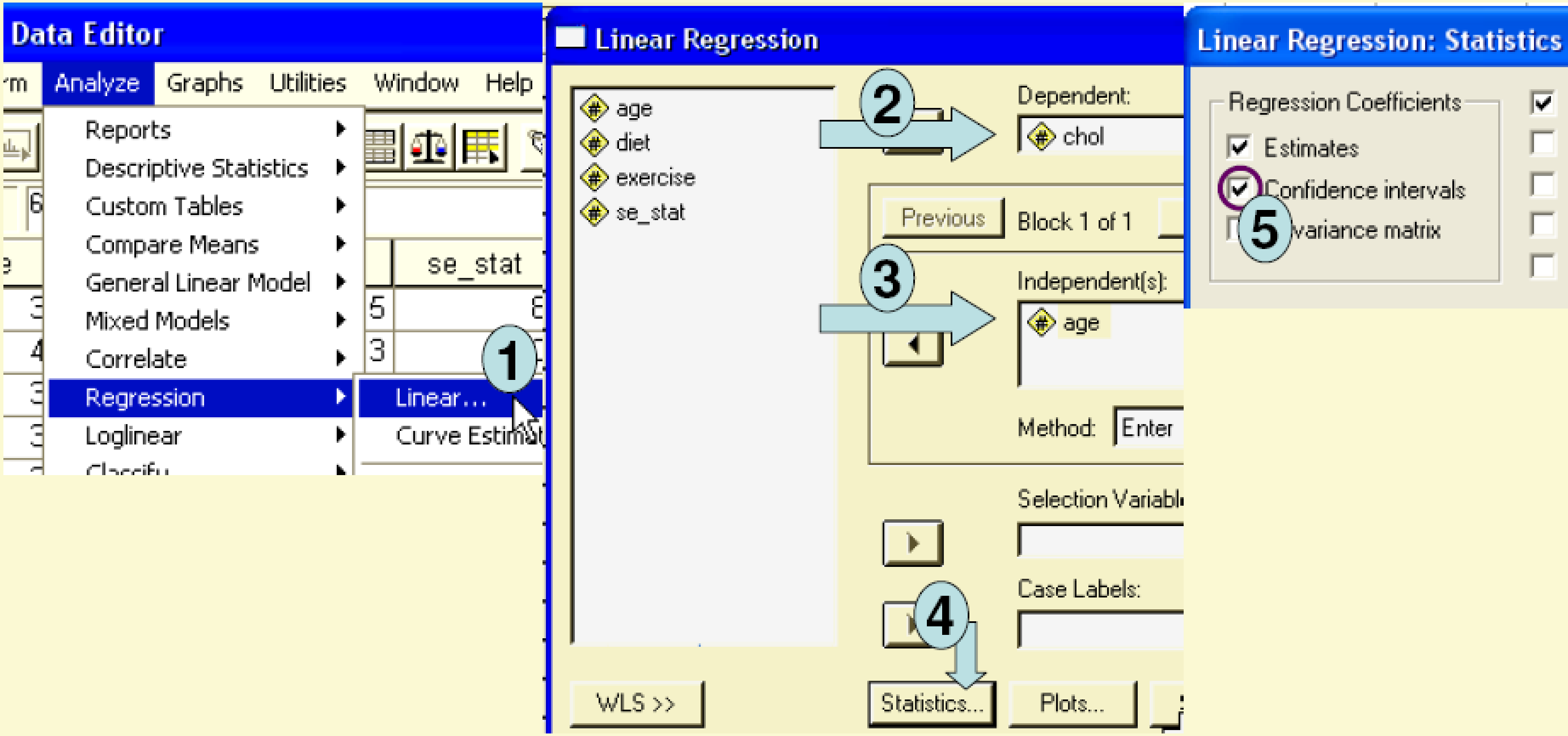
1) **To check the 'gross' relationship between dependent and each independent variable**
2) **Later this result will be compared with multiple linear regression result. This comparison indicates the confounding effects if it is present.**







Step 2: Simple Linear Regression



Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	*P* value Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B Upper Bound
1	(Constant)	5.895	.735		8.026	.000	4.434	7.357
	AGE age in year	5.776E-02	.018	.331	3.134	.002	.021	.094

a. Dependent Variable: CHOL cholesterol in mmo

Slope (*b*) = 0.058 (95% CI: .021, .094)

Table 3: Factors associated with blood cholesterol level (mmol/L) among the study population (n=82) using simple linear regression

Independent Variable	SLR[a]	
	b (95%CI)	*P* value
Age (year)	0.06 (0.02, 0.09)	0.002
Duration of exercise (hrs/wk)	- 0.62 (- 0.79, - 0.46)	<0.001
Diet inventory score	0.45 (0.30, 0.61)	<0.001
Socio-economic index	0.21 (0.17, 0.25)	<0.001

[a] Simple linear regression (Outcome as Cholesterol mmol/L)
b = crude regression coefficient

Regression Line

- In a scatterplot showing the association between 2 variables, the regression line is the “best-fit” line and has the formula

$y=a + bx$

a=place where line crosses Y axis

b=slope of line (rise/run)

Thus, given a value of X, we can predict a value of Y

Linear Regression

- Come up with a **Linear Regression Model** to predict a continous outcome with a continuous risk factor, i.e. predict BP with age. Usually LR is the next step after correlation is found to be strongly significant.
- $y = a + bx$; $a = y - bx$
  - e.g. BP = **constant (a) + regression coefficient (b) * age**
- b=

$$b = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

Example

$$b = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

$\sum x = 6426$ $\sum x^2 = 1338088$

$\sum y = 4631$ $\sum xy = 929701$

$n = 32$

b = (929701-(6426*4631/32))/(1338088-(6426^2/32)) = -0.00549

Mean x = 6426/32=200.8125

mean y = 4631/32=144.71875

y = a + bx

a = y – bx (replace the x, y & b value)

a = 144.71875+(0.00549*200.8125) = 145.8212106

Systolic BP = 145.82121 - 0.00549.chol

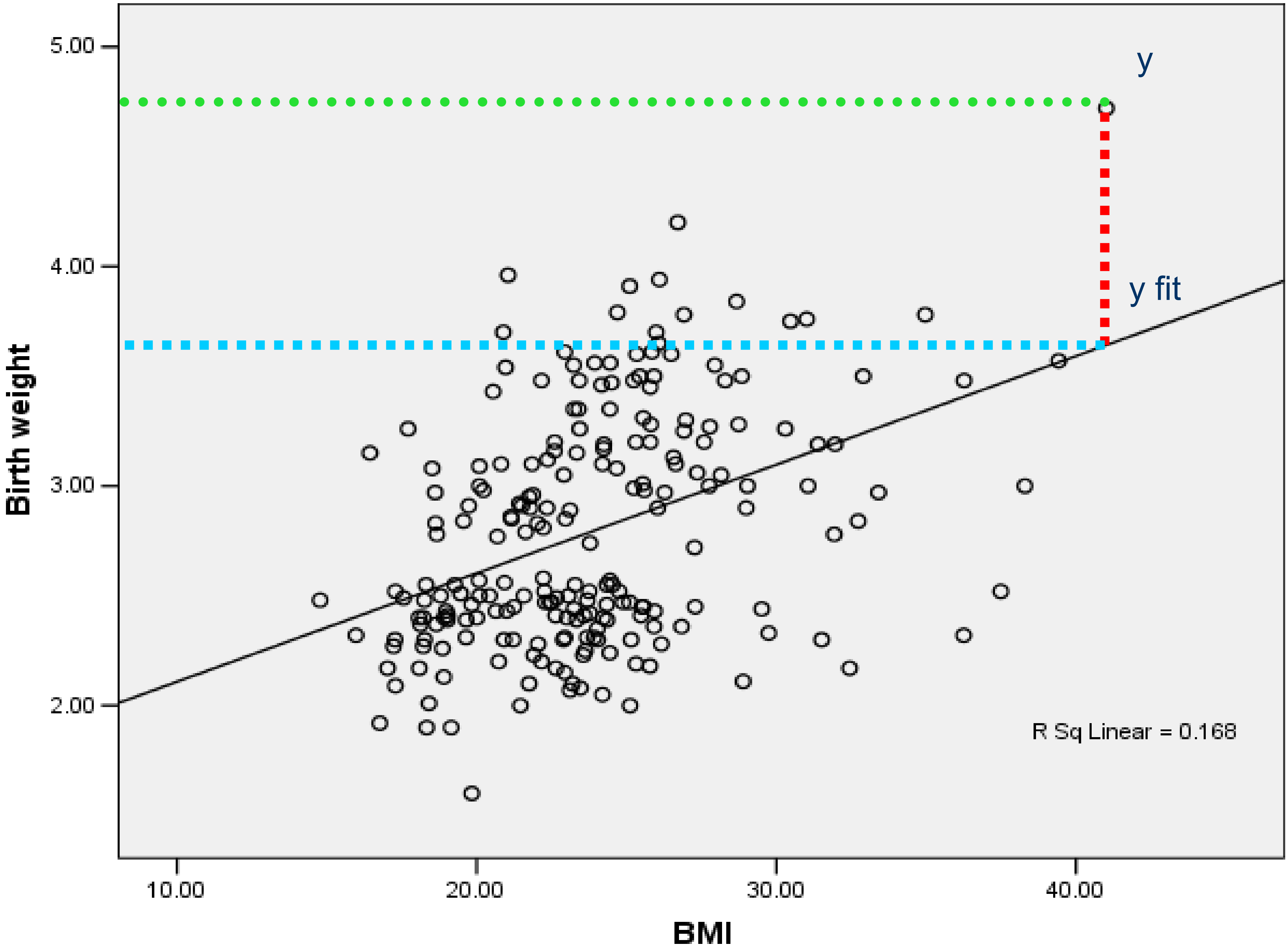
nores	chol x	bps1 y	x2	y2	xy
234	162	118	26244	13924	19116
235	210	126	44100	15876	26460
238	239	105	57121	11025	25095
240	187	112	34969	12544	20944
243	181	99	32761	9801	17919
244	180	99	32400	9801	17820
245	156	110	24336	12100	17160
274	191	133	36481	17689	25403
248	203	134	41209	17956	27202
253	169	129	28561	16641	21801
255	221	140	48841	19600	30940
256	223	117	49729	13689	26091
259	269	137	72361	18769	36853
231	151	164	22801	26896	24764
232	151	164	22801	26896	24764
233	249	164	62001	26896	40836
236	206	156	42436	24336	32136
237	252	147	63504	21609	37044
239	219	186	47961	34596	40734
241	129	170	16641	28900	21930
242	150	170	22500	28900	25500
246	194	176	37636	30976	34144
247	164	186	26896	34596	30504
249	223	157	49729	24649	35011
250	264	142	69696	20164	37488
251	232	159	53824	25281	36888
252	165	144	27225	20736	23760
254	232	155	53824	24025	35960
257	286	162	81796	26244	46332
258	180	151	32400	22801	27180
260	198	164	39204	26896	32472
261	190	155	36100	24025	29450
	6426	4631	1338088	688837	929701

Testing for significance

test whether the slope is significantly different from zero by:

$$t = b/SE(b)$$

$$SE_{(b)} = \frac{S_{res}}{\sqrt{\Sigma(x - \bar{x})^2}} \qquad S_{res} = \sqrt{\frac{\Sigma(y - y_{fit})^2}{n - 2}}$$



y
y fit
5.00
4.00
3.00
2.00
Birth weight
R Sq Linear = 0.168
10.00
20.00
30.00
40.00
BMI

	index	BMI	birth wgt	yfit	ytolakyfit	var
1	1	32.44	2.17	3.20	1.07	
2	2	20.74	2.20	2.63	.19	
3	3	22.04	2.28	2.70	.17	
4	4	14.77	2.48	2.34	.02	
5	5	18.33	1.90	2.51	.38	
6	6	19.03	2.41	2.55	.02	
7	7	27.29	2.45	2.95	.25	
8	8	21.00	2.43	2.64	.05	
9	9	18.92	2.40	2.54	.02	

nores	chol	bps1	X2	Y2	XY	Yfit	(Y-Yfit)^2	(X-mean)^2
1	162	118	26244	13924	19116	143.8294	667.1564	1513.70
2	210	126	44100	15876	26460	143.5659	308.5591	82.70
3	239	105	57121	11025	25095	143.4066	1475.07	1451.13
4	187	112	34969	12544	20944	143.6921	1004.39	193.38
5	181	99	32761	9801	17919	143.7251	2000.331	396.26
6	180	99	32400	9801	17820	143.7306	2000.822	437.07
7	156	110	24336	12100	17160	143.8623	1146.656	2016.57
8	191	133	36481	17689	25403	143.6702	113.8523	98.13
9	203	134	41209	17956	27202	143.6043	92.24219	4.38
10	169	129	28561	16641	21801	143.7909	218.7719	1018.01
11	221	140	48841	19600	30940	143.5055	12.28825	403.76
12	223	117	49729	13689	26091	143.4945	701.9575	488.13
13	269	137	72361	18769	36853	143.2419	38.96181	4636.76
14	151	164	22801	26896	24764	143.8898	404.4218	2490.63
15	151	164	22801	26896	24764	143.8898	404.4218	2490.63
16	249	164	62001	26896	40836	143.3517	426.3506	2313.01
17	206	156	42436	24336	32136	143.5878	154.0625	25.95
18	252	147	63504	21609	37044	143.3353	13.43025	2610.57
19	219	186	47961	34596	40734	143.5164	1804.853	327.38
20	129	170	16641	28900	21930	144.0105	675.452	5170.51
21	150	170	22500	28900	25500	143.8953	681.458	2591.45
22	194	176	37636	30976	34144	143.6537	1046.284	47.70
23	164	186	26896	34596	30504	143.8184	1779.288	1362.07
24	223	157	49729	24649	35011	143.4945	182.3991	488.13
25	264	142	69696	20164	37488	143.2694	1.611351	3980.82
26	232	159	53824	25281	36888	143.4451	241.9558	966.82
27	165	144	27225	20736	23760	143.8129	0.035006	1289.26
28	232	155	53824	24025	35960	143.4451	133.5164	966.82
29	289	162	83521	26244	46818	143.1321	355.9961	7760.51
30	180	151	32400	22801	27180	143.7306	52.8449	437.07
31	198	164	39204	26896	32472	143.6317	414.8664	8.45
32	190	155	36100	24025	29450	143.6757	128.2409	118.95
Mean	200.91						18682.55	48186.72

$$S_{res} = \sqrt{\frac{\Sigma(y - y_{fit})^2}{n-2}}$$

Sres = / 48186.72

√ 30

Sres = 40.077724

$$SE_{(b)} = \frac{S_{res}}{\sqrt{\Sigma(x-\bar{x})^2}}$$

SE(b) = 40.077724

√ 18682.55

SE(b) = 0.29321421

t = b/SE(b)

t = 0.00549

0.29321421

t = 0.01872351

p > 0.05

Conclusion; Regression coefficient (b) of 0.00549 (slope) was not significantly different than 0.

Coefficients[a]

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	145.594	23.237		6.265	<.001
	chol	-.004	.114	-.007	-.038	.970

a. Dependent Variable: bps1

$b = (929701-(6426*4631/32))/(1338088-(6426^2/32)) = -0.00549$

Mean x = 6426/32=200.8125
mean y = 4631/32=144.71875

y = a + bx

a = y – bx (replace the x, y & b value)

a = 144.71875+(0.00549*200.8125)
= 145.8212106

Systolic BP = 145.82121 - 0.00549.chol

$S_{res} = (48186.72/30)^{0.5}$
$= 40.077724$

$SE(b) = \sqrt{\dfrac{40.07772397}{18682.55}}$

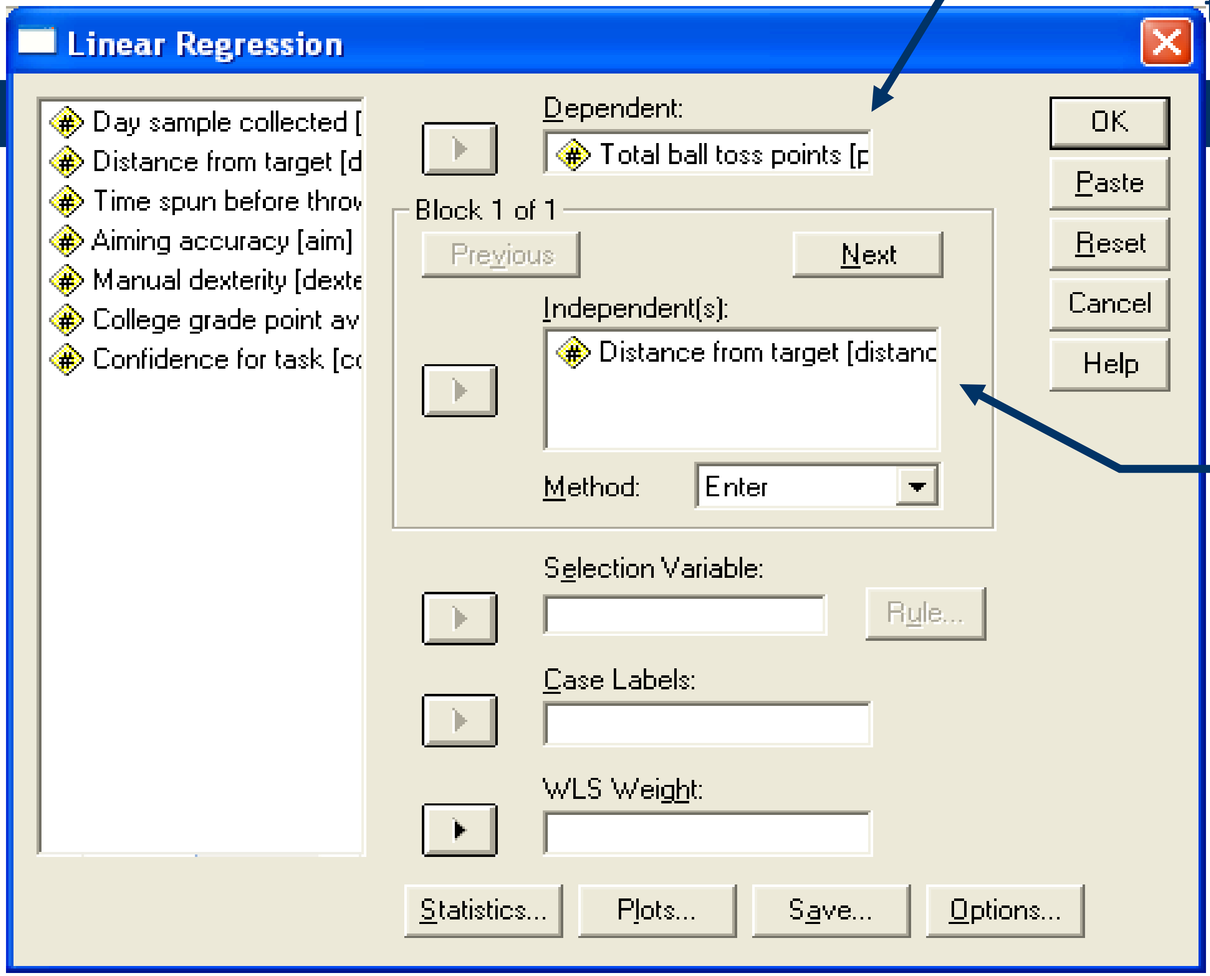
SE(b) = 0.293214208

$t = \dfrac{0.00549}{0.293214208}$

t = 0.018723513

$p > 0.05$

SPSS Regression Set-up



- “**Criterion**,”
- y-axis variable,
- what you’re trying to predict

- “**Predictor**,”
- x-axis variable,
- what you’re basing the prediction on

Getting Regression Info from SPSS

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.777[a]	.603	.581	18.476

a. Predictors: (Constant), Distance from target

$y' = a + b\,(\mathbf{x})$

$y' = 125.401 - 4.263(\mathbf{20})$

a

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	125.401	14.265		8.791	.000
	Distance from target	-4.263	.815	-.777	-5.230	.000

a. Dependent Variable: Total ball toss points

b

Birthweight=1.615+0.049mBMI

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.615	.181		8.909	.000
	BMI	.049	.007	.410	6.605	.000

a. Dependent Variable: Birth weight