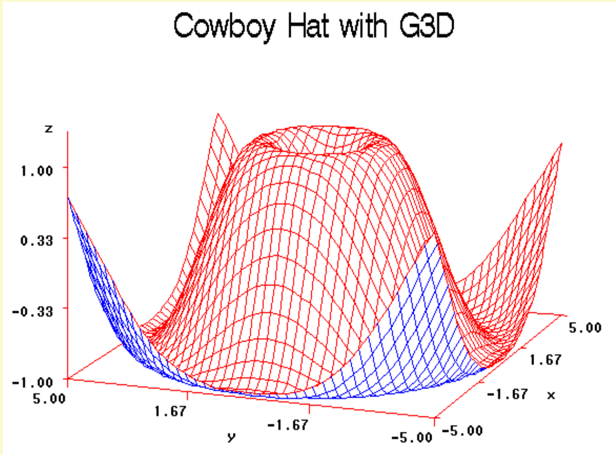
L4: Multiple Linear Regression

Presented by

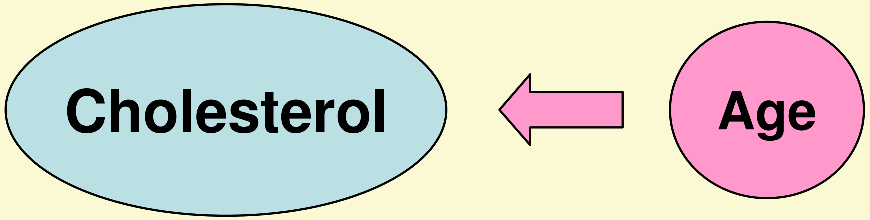
Dr Mohd. Ayub Sadiq @ Lin Naing
MBBS, MPH, MHltSc(OH), MMedStat.
Institute of Medicine
Universiti Brunei Darussalam



Contents

Simple Linear Regression

- **To determine the relationship between age and blood cholesterol level**



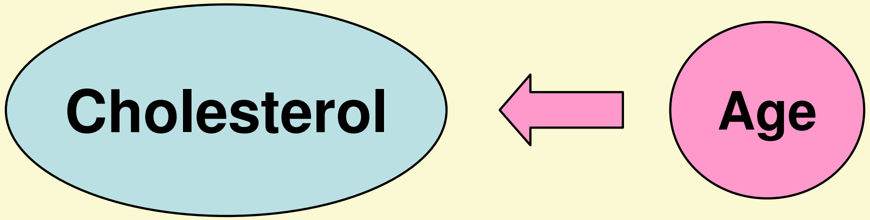
► **Here, we may use either '<u>correlation analysis</u>' or '<u>regression analysis</u>', as both cholesterol and age are numerical variables.**

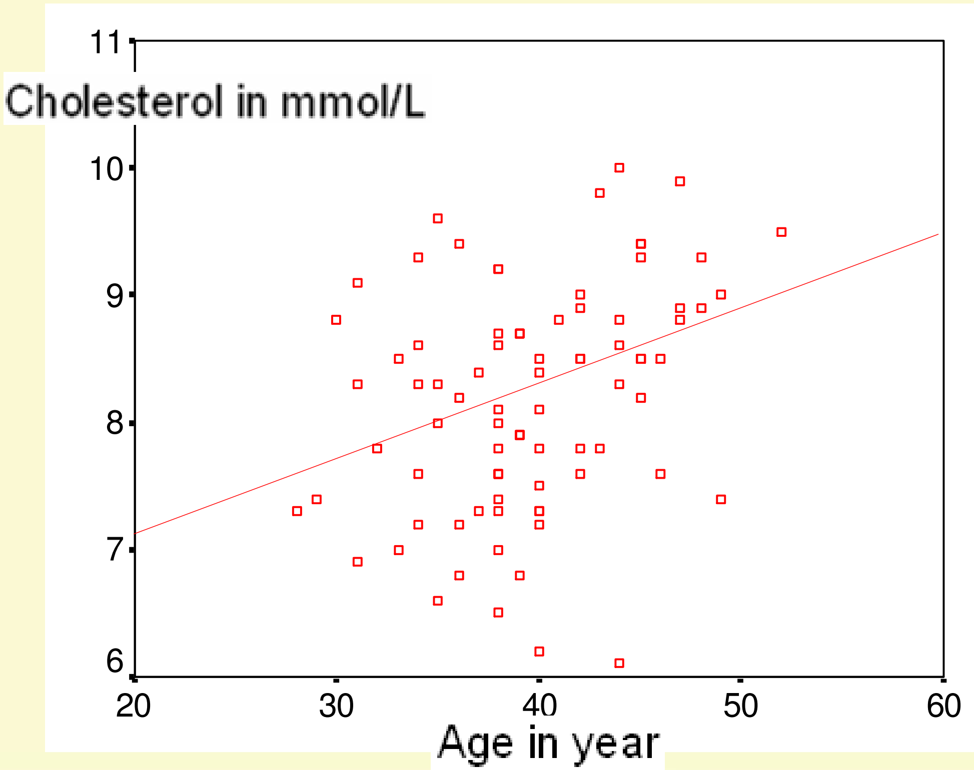
► ***Correlation* can give the strength of relationship, but *regression* can describe the relationship in more detail.**

► **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.**

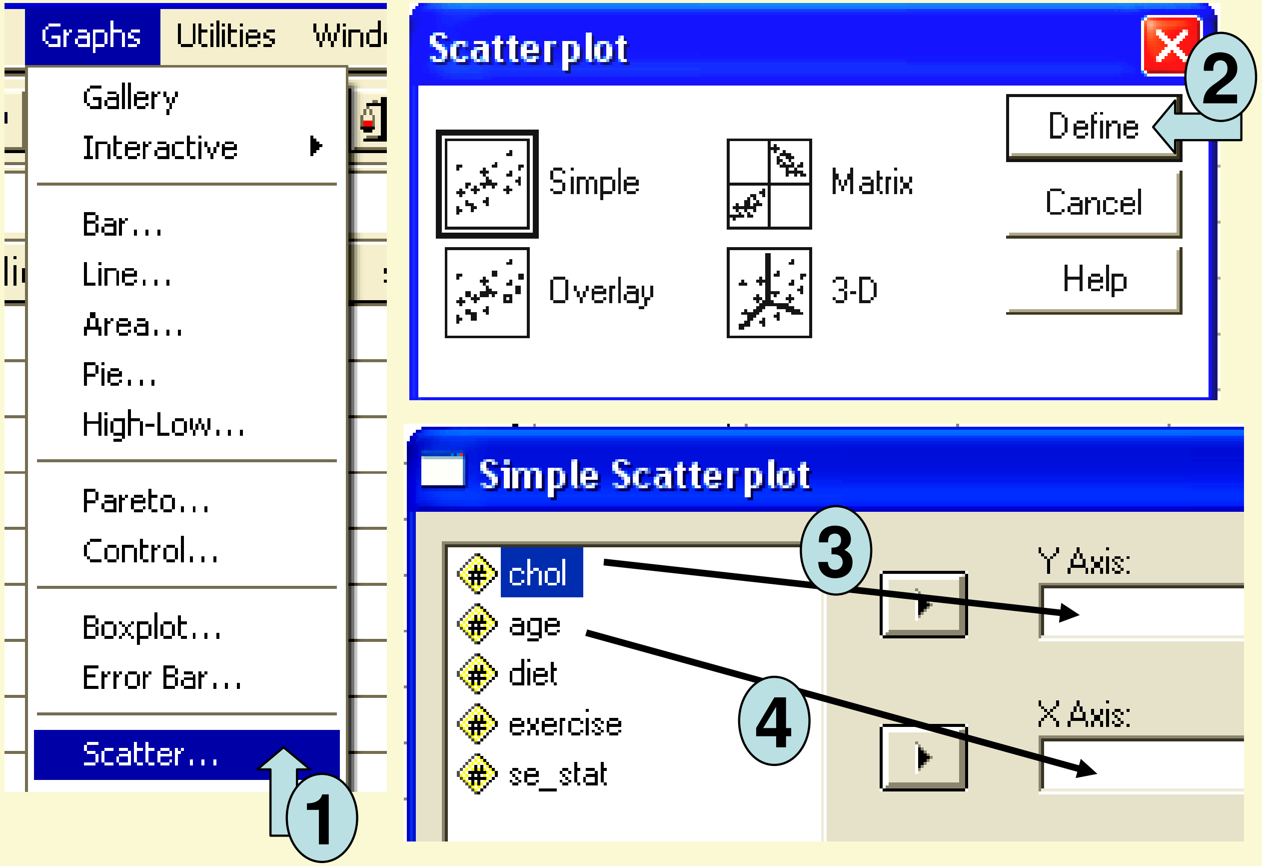
Simple Linear Regression

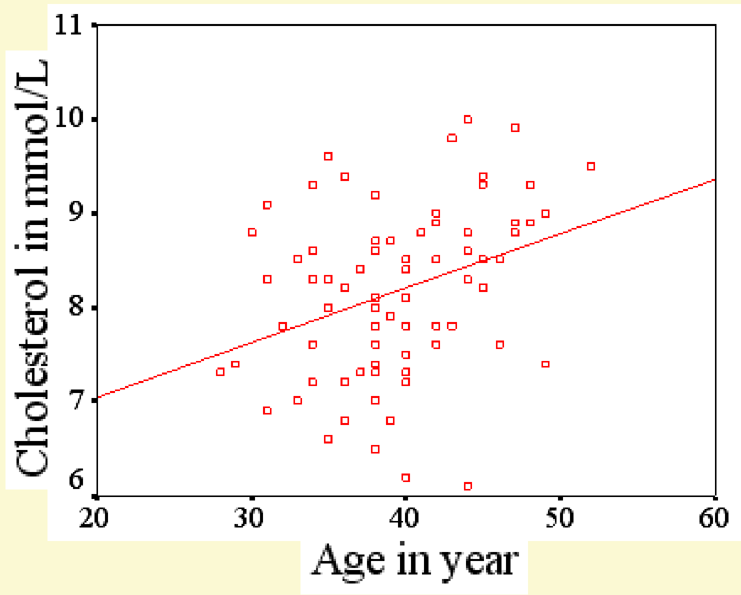
- **To determine the relationship between age and blood cholesterol level**



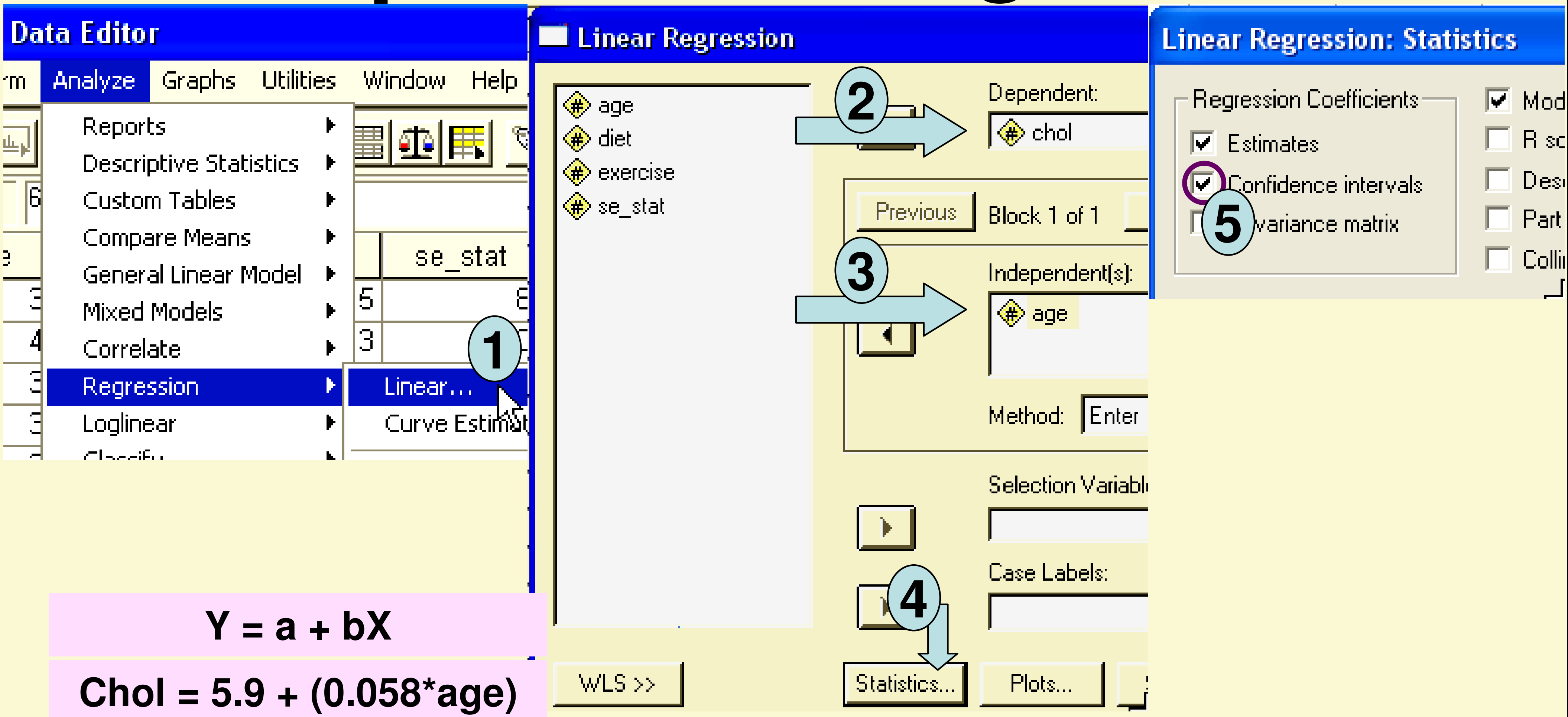


Simple Linear Regression





Simple Linear Regression



Y = a + bX

Chol = 5.9 + (0.058*age)

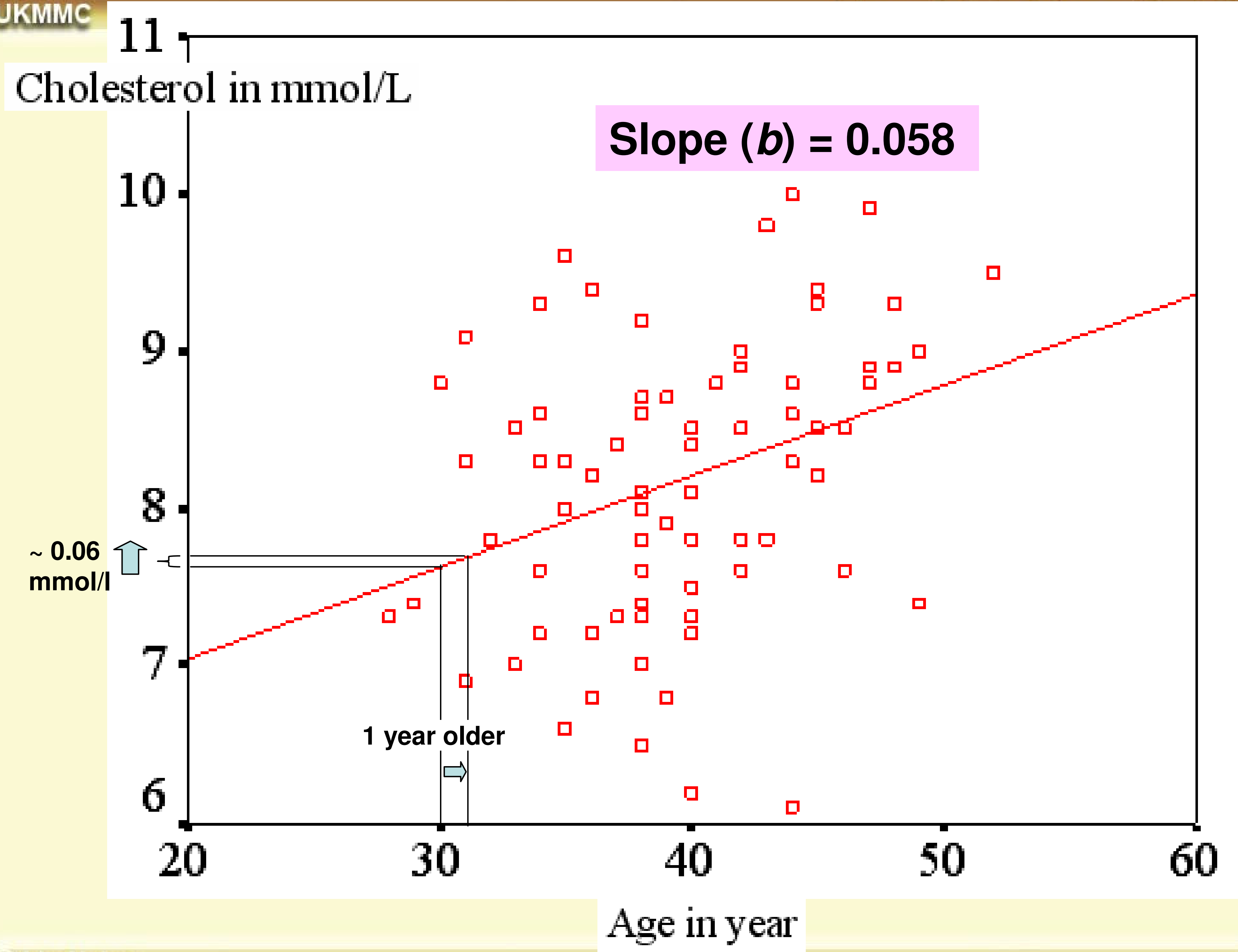
Coefficients[a]

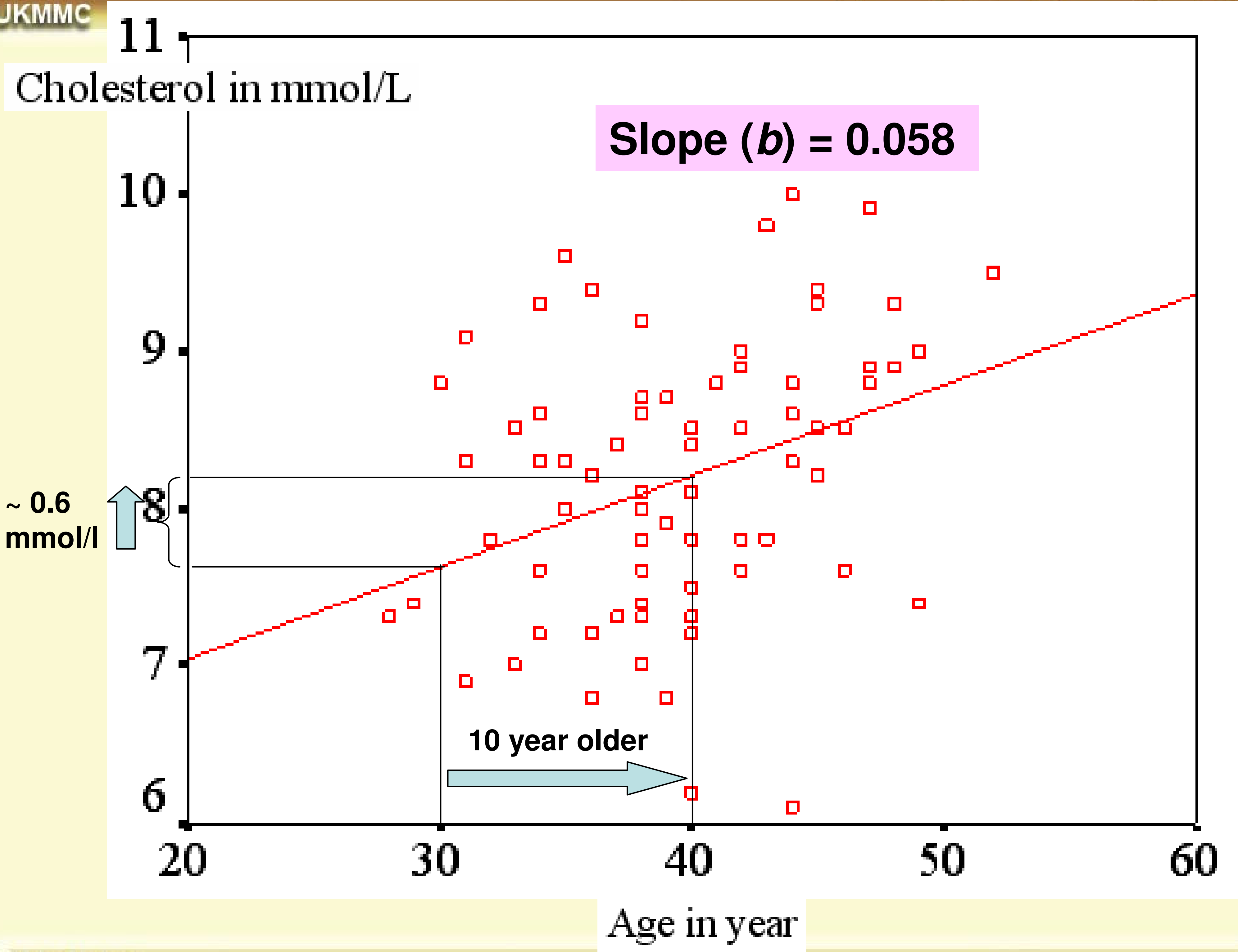
Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	***P* value** Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B Upper Bound
1	(Constant)	5.895	.735		8.026	.000	4.434	7.357
	AGE age in year	5.776E-02	.018	.331	3.134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

a. Dependent Variable: CHOL cholesterol in mmol/L

H_0: β=0

Slope (*b*) = 0.058 (95% CI: 0.021, 0.094)


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

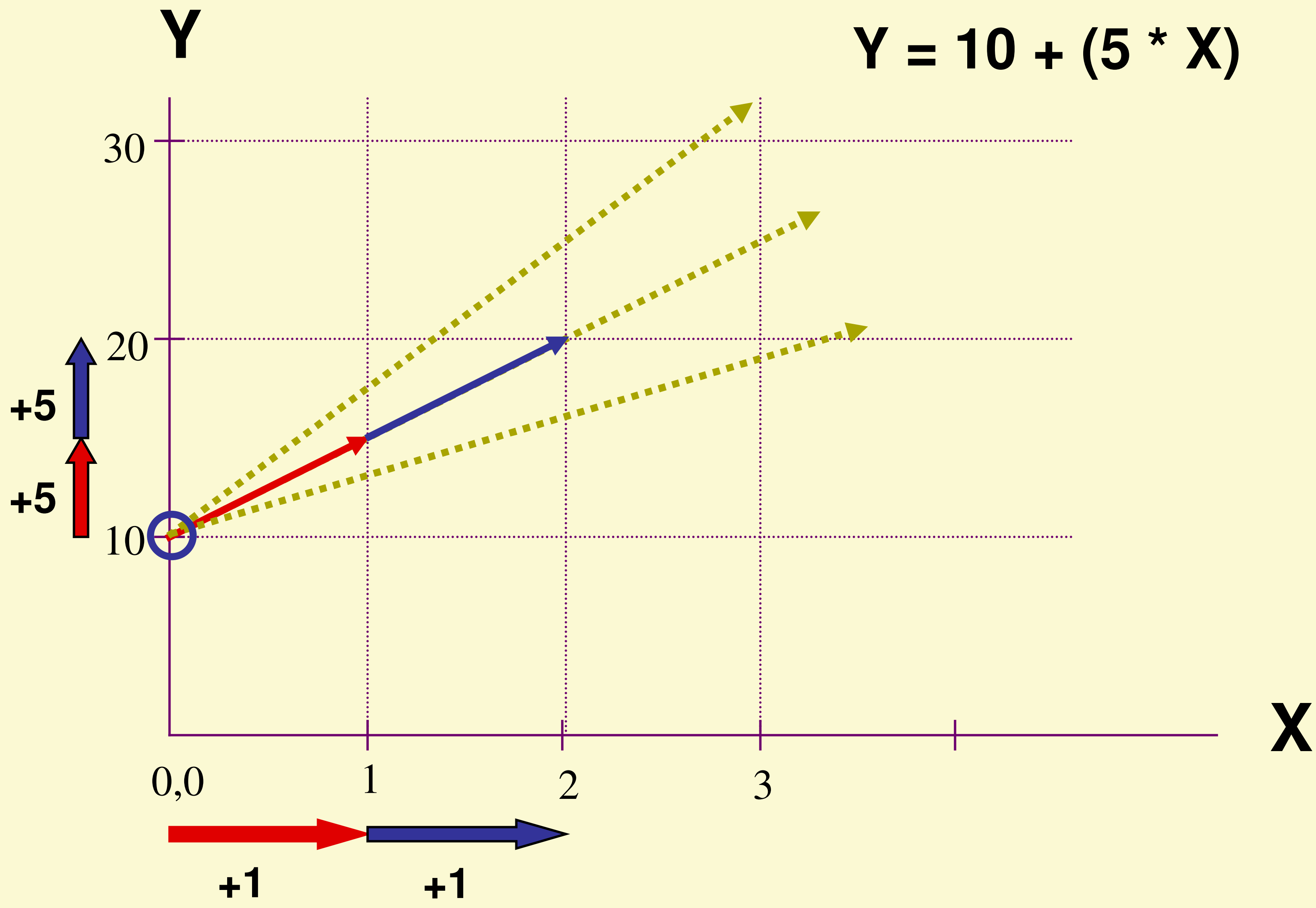

Cholesterol in mmol/L
11
10
9
8
7
6
Slope (b) = 0.058
~ 0.6 mmol/l
10 year older
20
30
40
50
60
Age in year

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

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

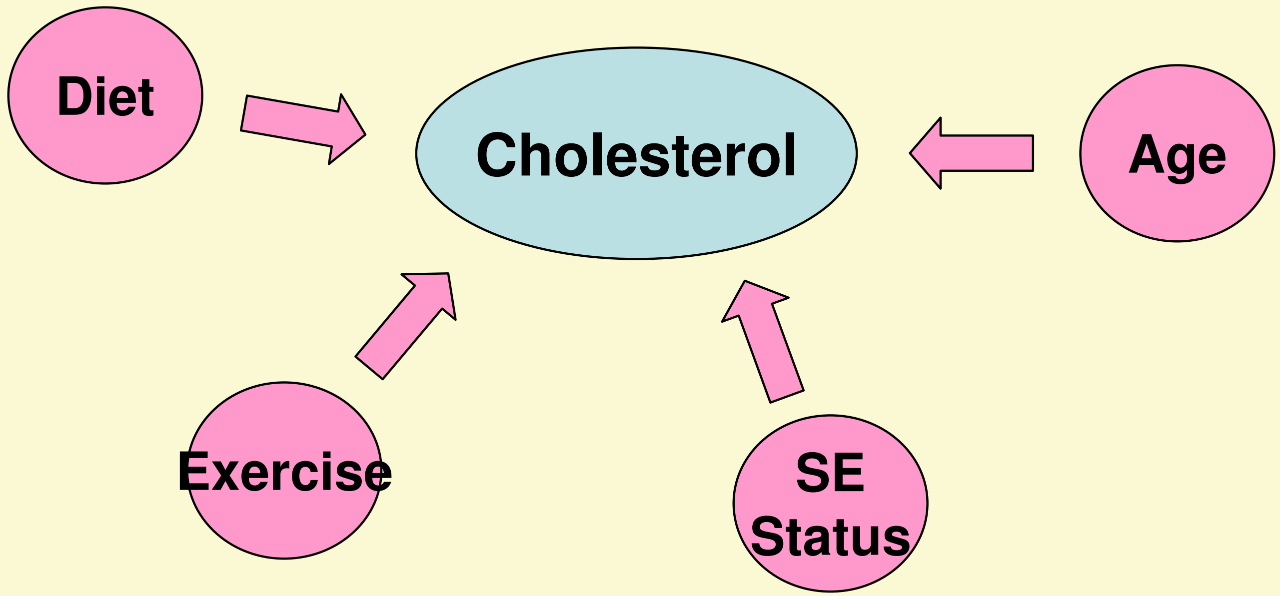
$$Y = Constant + (slope * X)$$

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

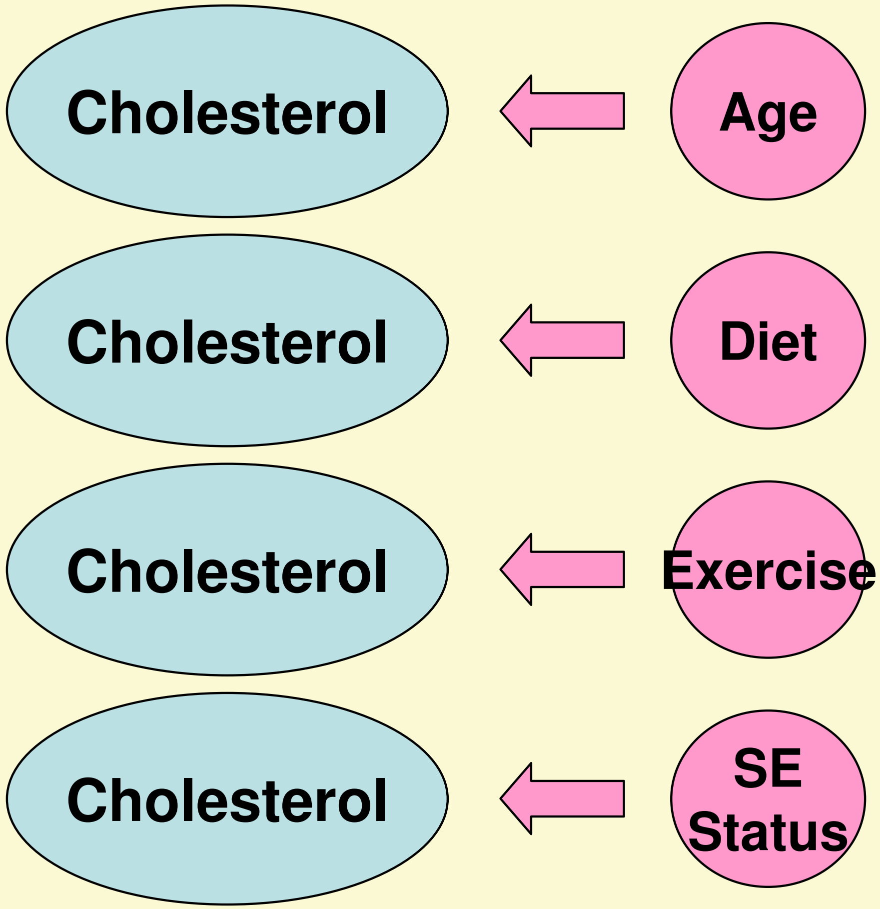


Basic Theory of MLR

- Most of the outcomes (events) are determined (influenced) by more than one factors (e.g. blood pressure, cholesterol level, etc.)

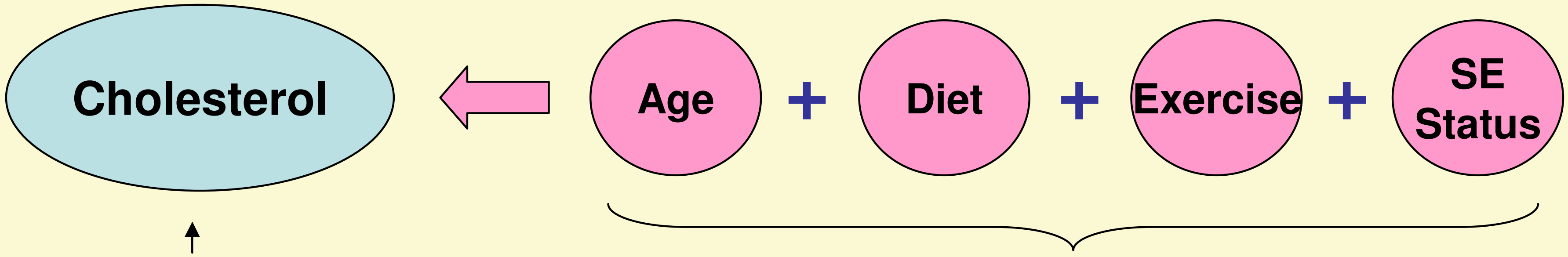


Basic Theory



- If we look at each factor to the outcome at one time, it will not be realistic.
- We should look at the relationship of these factors to the outcome at the same time.

Basic Theory



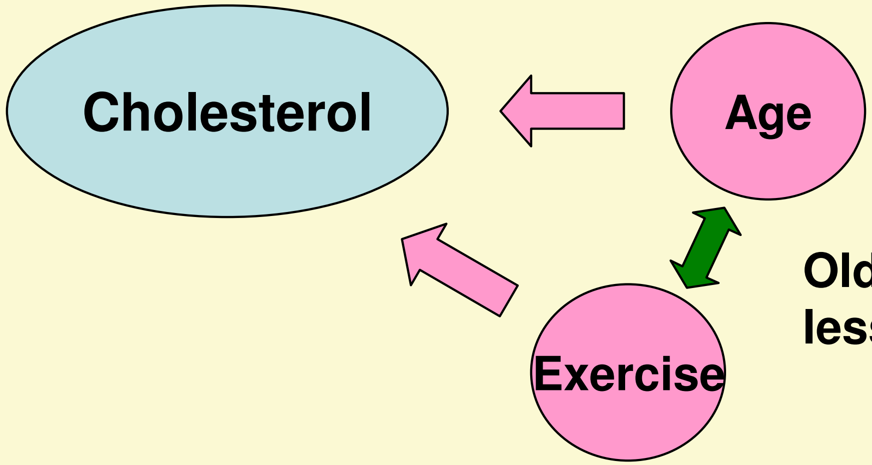
Dependent variable or
Outcome variable

Independent variables or
Explanatory variables

When we look at the relation of these factors (explanatory variables) to the outcome at the same time, ….

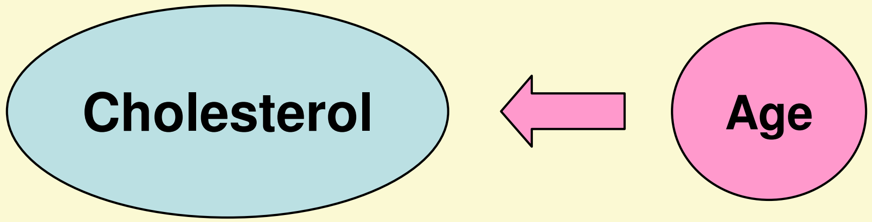
- **We will obtain the "independent effect" of explanatory variables to outcome.**
- **We can also study the "interaction" (IA) between independent variables (Synergistic/Antagonistic IA).**

Independent Effect / Confounding



Older people have
less exercise.

Independent Effect / Confounding



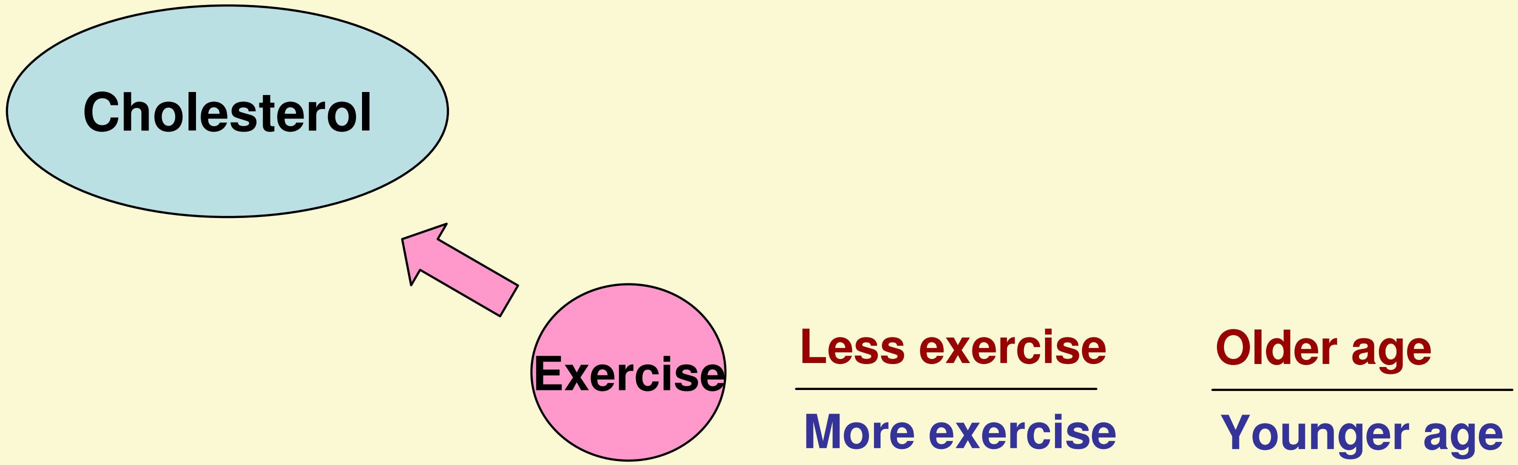
Older people / Younger people

Less exercise / More exercise

Effect that we found here, is not only the pure effect of age, but also additional effect from exercise. (Older people have less exercise – so that the relationship of being higher cholesterol among older age is exaggerated by the effect of less exercise).

In this example, the result (of the relationship between cholesterol and age) is confounded by exercise.

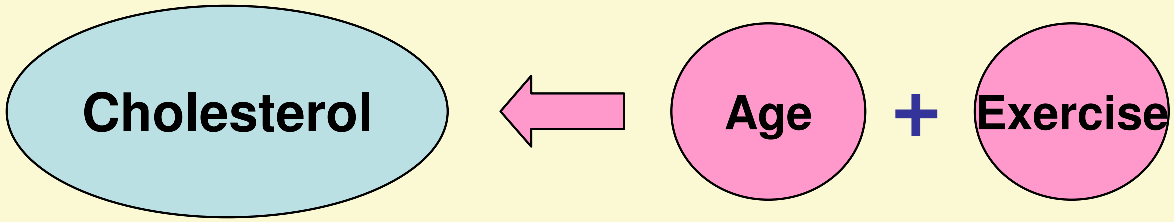
Independent Effect / Confounding



Effect that we found here, is not only the pure effect of exercise, but also additional effect from age. (Less exercise people are older people – so that the relationship of being higher cholesterol among less exercise people is exaggerated by the effect of older age).

In this example, the result (of the relationship between cholesterol and exercise) is confounded by age.

Independent Effect / Confounding



But, if we subject them together in the regression model, the confounding effects were eliminated and we can get the “independent effect” of each independent variable.

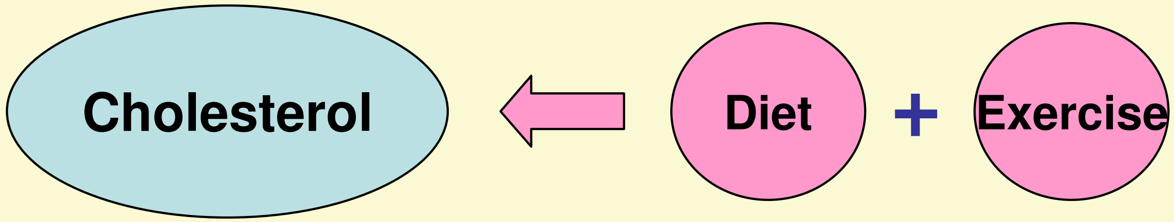
Interaction

Cholesterol ← Diet + Exercise

1 mmol/l ↑	←	1 unit ↑ + no change	Diet Effect
1 mmol/l ↑	←	no change + 1 hr/wk ↓	Exerc. Effect
2 mmol/l ↑	←	1 unit ↑ + 1 hr/wk ↓	Combined Effect (No IA)
2.5 mmol/l ↑	←	1 unit ↑ + 1 hr/wk ↓	Combined Effect (Syn. IA)
1.5 mmol/l ↑	←	1 unit ↑ + 1 hr/wk ↓	Combined Effect (Ant. IA)

IA=Interaction; Syn. IA=Synergistic Interaction; Ant. IA= Antagonistic Interaction

Interaction



Example:

Those with higher cholesterol diet, their cholesterol level will be higher.

*Say, **1 unit more in cholesterol diet score**, cholesterol level will be higher for 1 mmol/L.*

Those with less exercise, their cholesterol level will be higher.

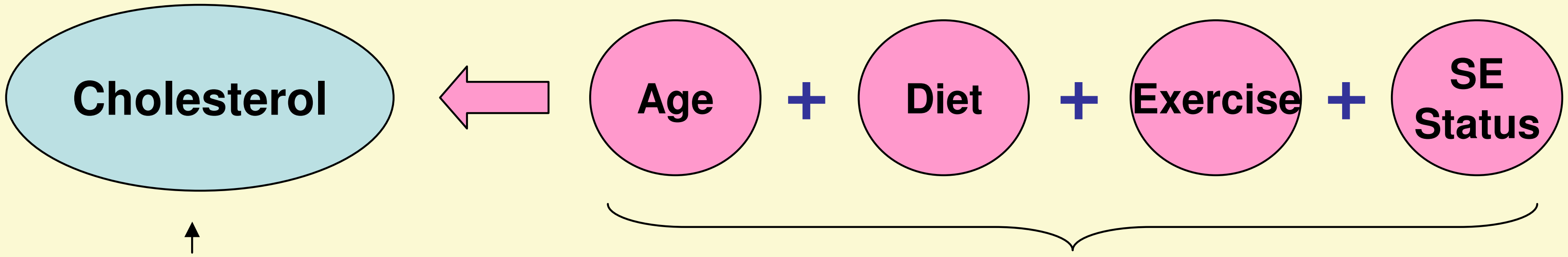
*Say, **1 hour less exercise in a week**, cholesterol will higher for 1 mmol/L.*

It means … for 1 unit more in cholesterol diet AND 1 hour less exercise in a week, there should be an increase in cholesterol for 2 mmol/L.

If it doesn't happen as above, but it increases for 3 mmol/L, it means that there is a synergistic interaction between diet and exercise.

If it doesn't happen as above, but it increases only for 1.5 mmol/L, it means that there is an antagonistic interaction between diet and exercise.

Basic Theory

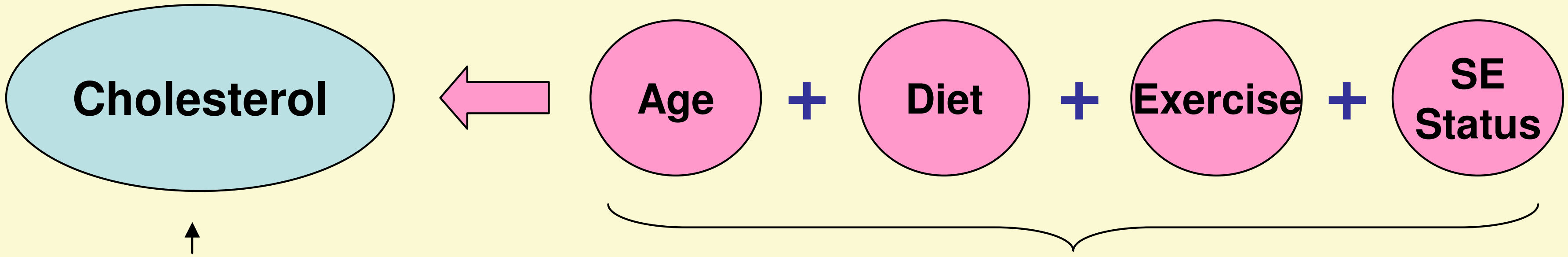


Dependent variable or
Outcome variable

Independent variables or
Explanatory variables

- **This analysis is used for ….**
  - **Exploring associated / influencing / risk factors to outcome (exploratory study)**
  - **Developing prediction model (exploratory study)**
  - **Confirming a specific relationship (confirmatory study)**

Basic Theory



Dependent variable or
Outcome variable

Numerical

Independent variables or
Explanatory variables

Numerical (MLR analysis)
Categorical or Mixed (GLR analysis)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots\ldots + \beta_n X_n$$

- **If the dependent variable is numerical and independent variables are numerical, it will be called Multiple Linear Regression (MLR) analysis.**
- **MLR can be with categorical independent variables, but special name is given as General Linear Regression analysis.**

Steps in Handling MLR

Step 1: Data exploration (Descriptive Statistics)

Step 2: Scatter plots and Simple Linear Regression

Step 3: Variable selection

⇨ Preliminary main-effect model

Step 4: Checking interaction & multicollinearity[a]

⇨ Preliminary final model

Step 5: Checking model assumptions & outliers[a]

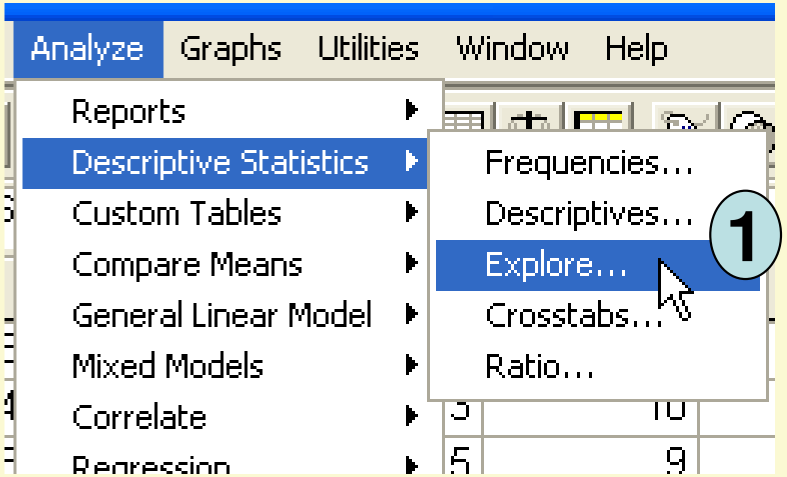
⇨ Final model

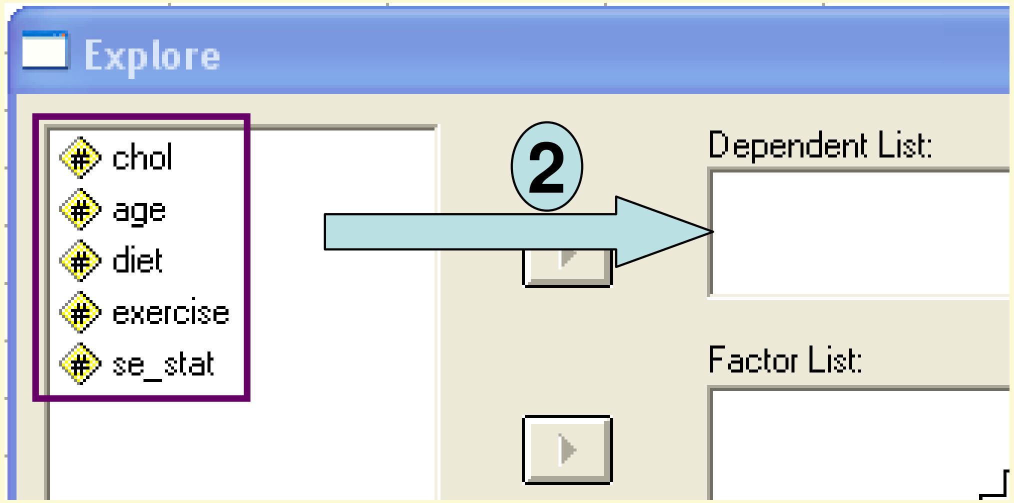
Step 6: Interpretation & data presentation

[a] need remedial measures if problems are detected

Neter J, Kutner MH, Nachtsheim CJ, Wasserman W. (1996). Applied linear statistical models (Fourth Ed.). Chicago: Irwin.

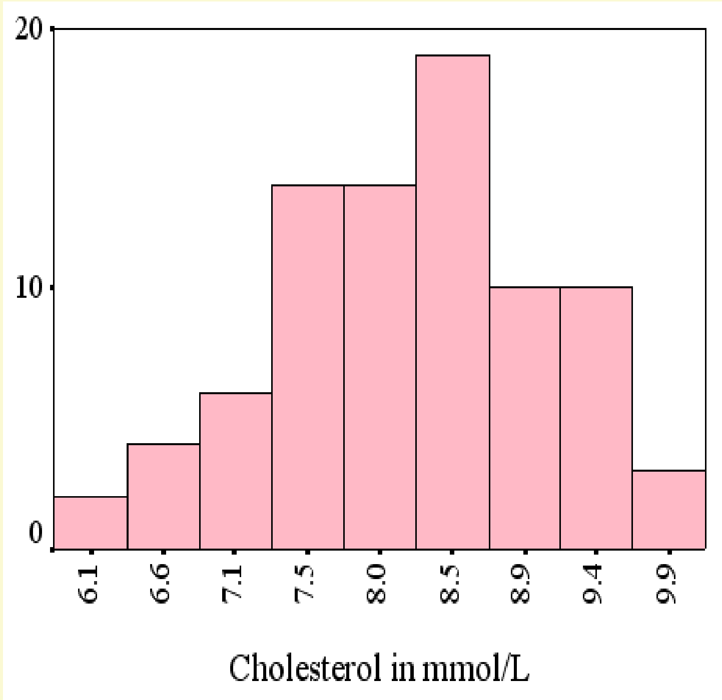
Step 1: Data Exploration





Descriptives

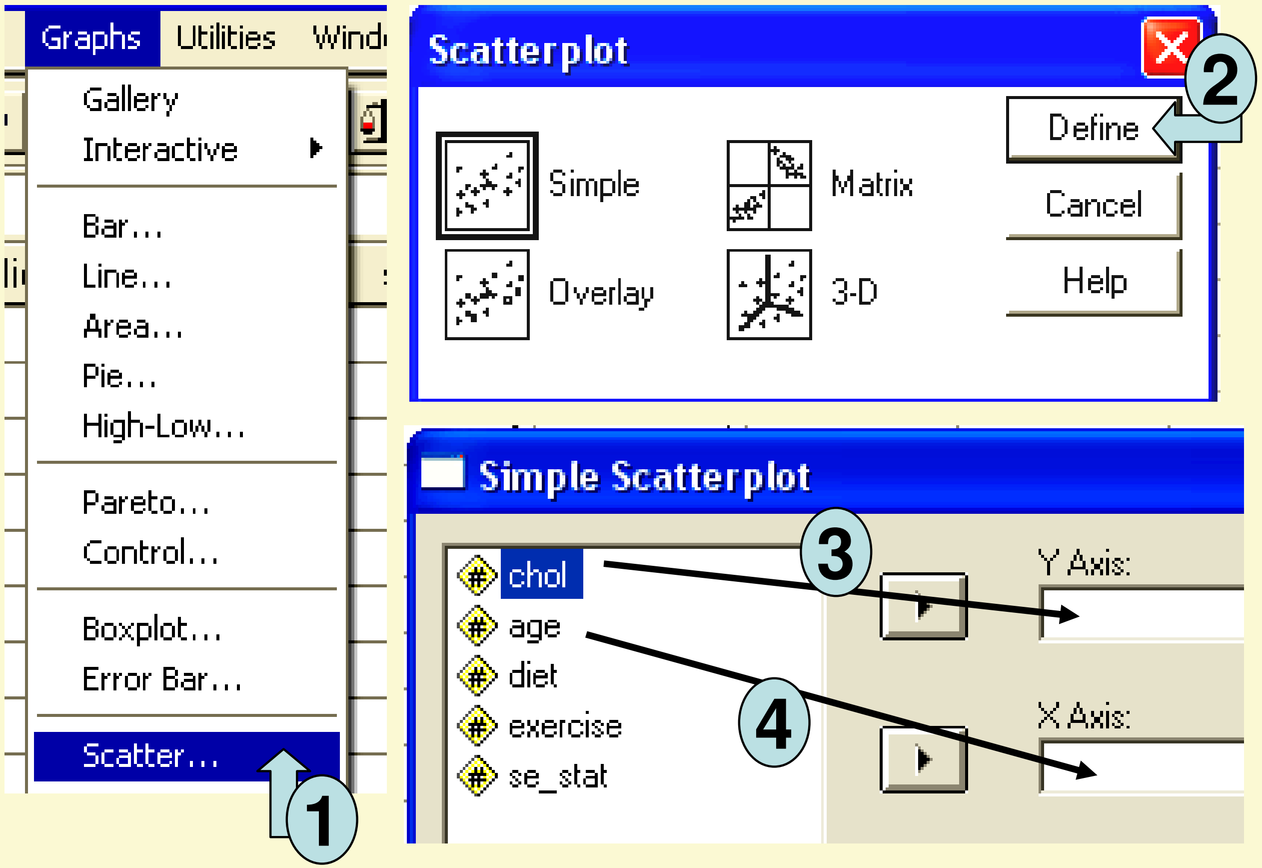
			Statistic
CHOL cholesterol in mmol/L	Mean		8.179
	95% Confidence Interval for Mean	Lower Bound	7.984
		Upper Bound	8.375
	5% Trimmed Mean		8.189
	Median		8.300
	Variance		.790
	Std. Deviation		.8890
	Minimum		6.1
	Maximum		10.0
	Range		3.9
	Interquartile Range		1.325
	Skewness		-.145
	Kurtosis		-.549

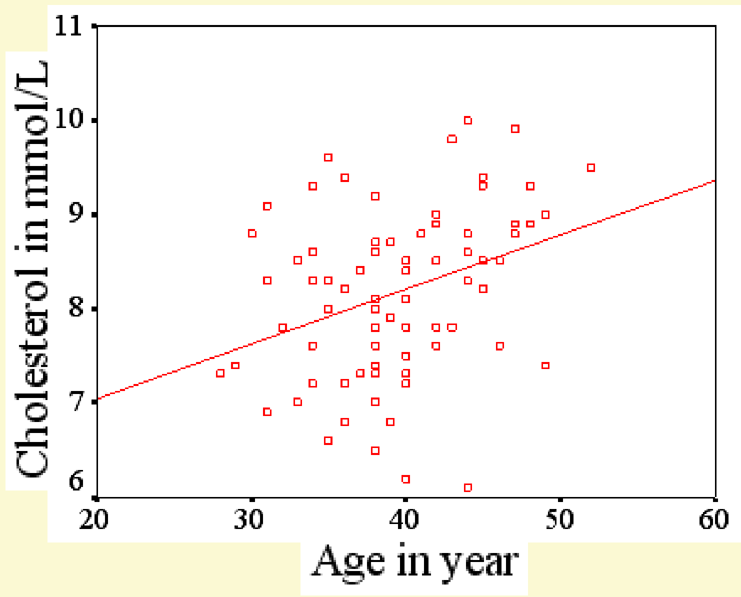


Step 2: Simple Linear Regression

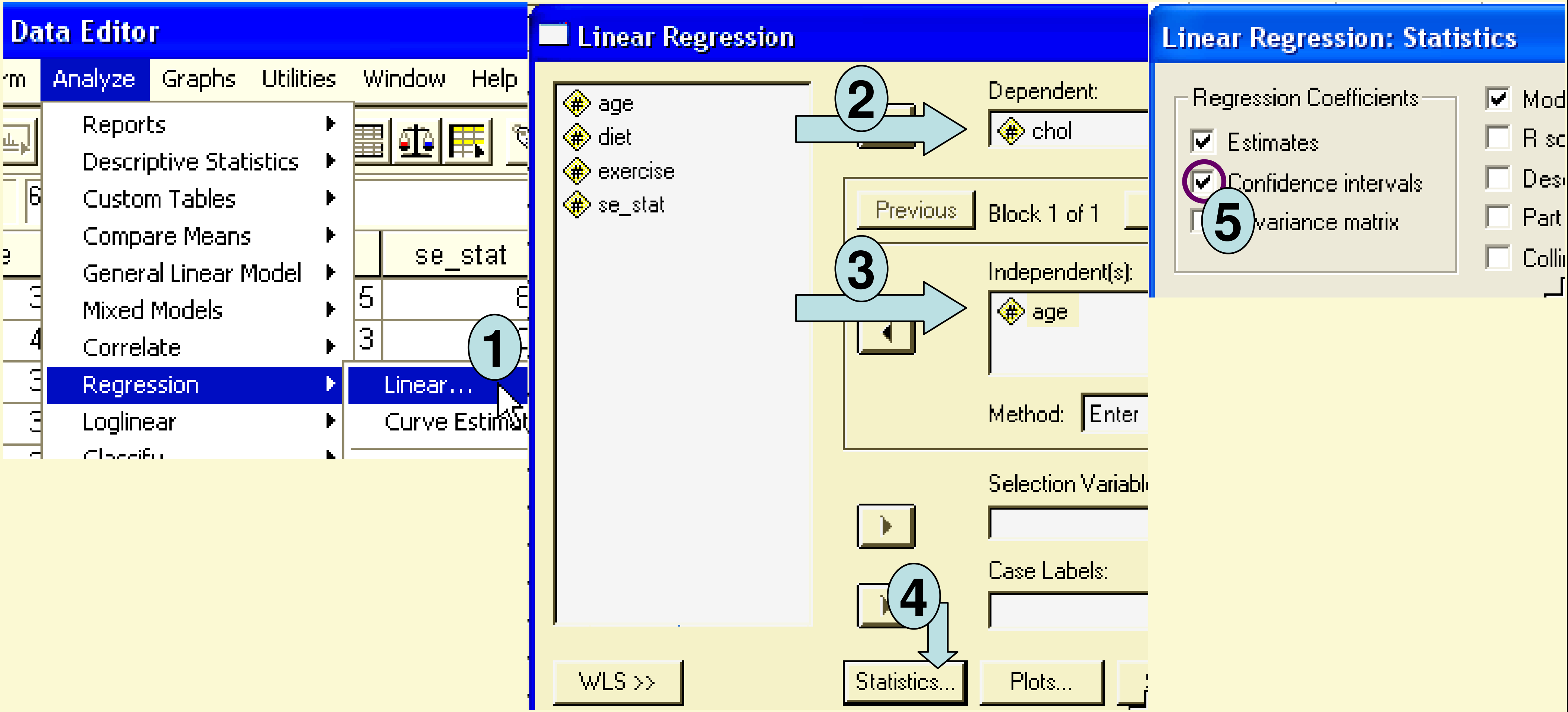
Two main reasons:

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	Sig. (*P* value)	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

Steps in Handling MLR

Step 1: Data exploration (Descriptive Statistics)

Step 2: Scatter plots and Simple Linear Regression

Step 3: Variable selection

⇨ Preliminary main-effect model

Step 4: Checking interaction & multicollinearity[a]

⇨ Preliminary final model

Step 5: Checking model assumptions & outliers[a]

⇨ Final model

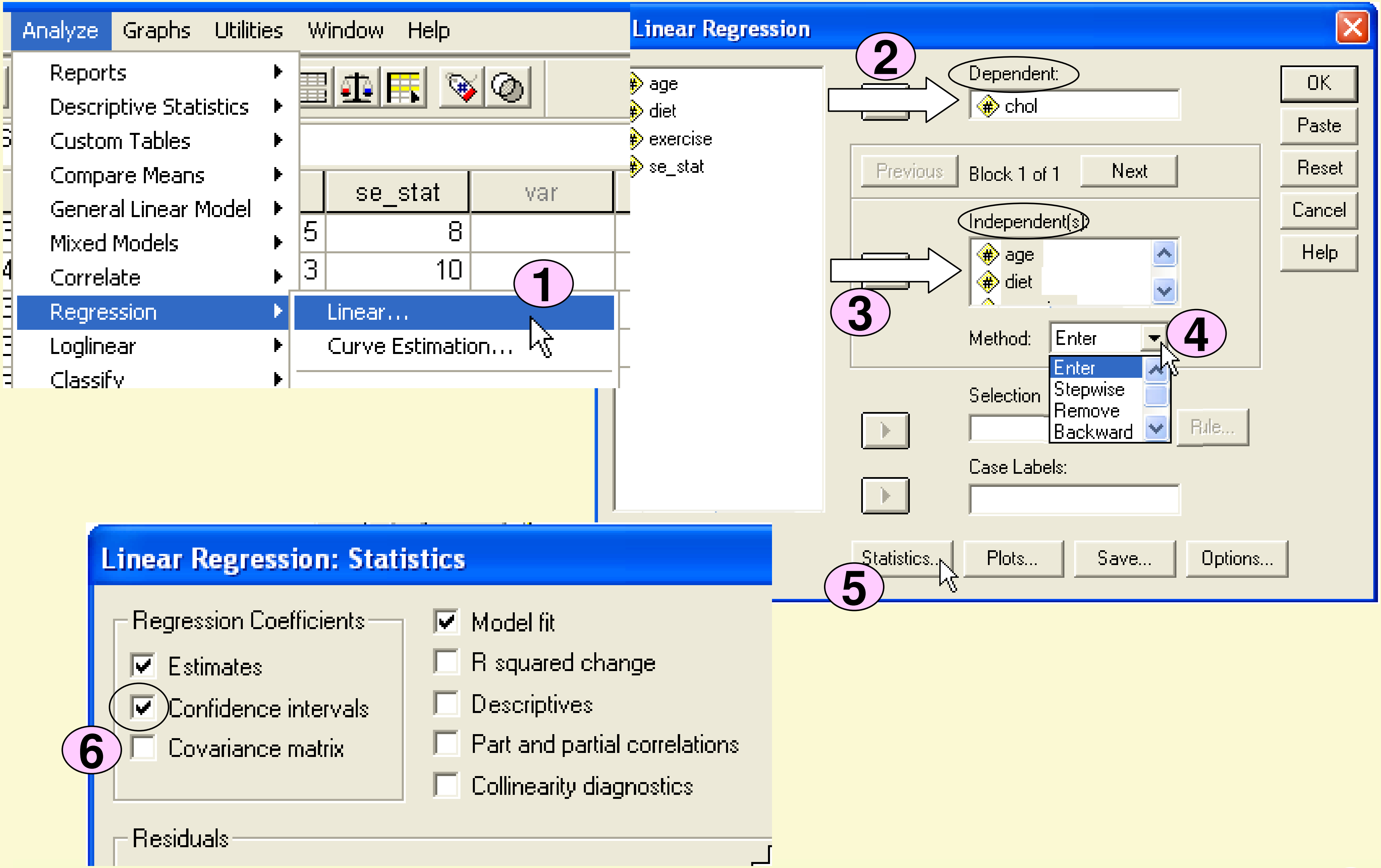
Step 6: Interpretation & data presentation

[a] need remedial measures if problems are detected

Step 3: Variable Selection

- Automatic / Manual methods
  - Forward method
  - Backward method
  - Stepwise method
  - All possible models method
- Nowadays, as computers are faster, automatic methods can be done easily.
- In SPSS, *forward*, *backward* and *stepwise* can be used.
- All 3 methods should be used for this step. Take the biggest model (all selected variables should be significant) for further analysis.

Step 3: Variable Selection



Result: Stepwise

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B Upper Bound
1	(Constant)	5.845	.241		24.264	.000	5.366	6.325
	socio-econmic status (index)	.211	.021	.748	10.069	.000	.170	.253
2	(Constant)	7.660	.587		13.048	.000	6.492	8.829
	socio-econmic status (index)	.158	.025	.559	6.235	.000	.108	.208
	duration of exercise (hours/week)	-.288	.086	-.301	-3.352	.001	-.460	-.117
3	(Constant)	8.593	.633		13.574	.000	7.332	9.853
	socio-econmic status (index)	1.369E-02	.052	.048	.262	.794	-.090	.118
	duration of exercise (hours/week)	-.550	.117	-.574	-4.688	.000	-.784	-.317
	diet inventory (higher the score, higher	.372	.120	.451	3.106	.003	.134	.610

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots\ldots + \beta_n X_n$$

Model		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
4								.516
	(hours/week)							-.450

Cholesterol = 7.297 – (0.540*exercise) + (0.394*diet) + (0.033*age)

Model		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
5	(Constant)	7.297	.620		11.763	.000	6.062	8.532
	duration of exercise (hours/week)	-.540	.062	-.563	-8.702	.000	-.663	-.416
	diet inventory (higher the score, higher cholesterol content)	.394	.052	.478	7.527	.000	.290	.498
	age in year	3.281E-02	.011	.188	2.914	.005	.010	.055

***P* values**

a. Dependent Variable: cholesterol in mmol/L

Result: Forward

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B Upper Bound
1	(Constant)	5.845	.241		24.264	.000	5.366	6.325
	socio-econmic status (index)	.211	.021	.748	10.069	.000	.170	.253
2	(Constant)	7.660	.587		13.048	.000	6.492	8.829
	socio-econmic status (index)	.158	.025	.559	6.235	.000	.108	.208
	duration of exercise (hours/week)	-.288	.086	-.301	-3.352	.001	-.460	-.117
3	(Constant)	8.593	.633		13.574	.000	7.332	9.853
	socio-econmic status (index)	1.369E-02	.052	.048	.262	.794	-.090	.118
	duration of exercise (hours/week)	-.550	.117	-.574	-4.688	.000	-.784	-.317
	diet inventory (higher the score, higher cholesterol content)	.372	.120	.451	3.106	.003	.134	.610
4	(Constant)	7.151	.783		9.131	.000	5.591	8.710
	socio-econmic status (index)	1.545E-02	.050	.055	.309	.758	-.084	.115
	duration of exercise (hours/week)	-.511	.113	-.533	-4.519	.000	-.736	-.286
	diet inventory (higher the score, higher cholesterol content)	.363	.114	.440	3.168	.002	.135	.591
	age in year	3.285E-02	.011	.188	2.900	.005	.010	.055

a. Dependent Variable: cholesterol in mmol/L

***P* values**

Result: Backward

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B Upper Bound
1	(Constant)	7.151	.783		9.131	.000	5.591	8.710
	age in year	3.285E-02	.011	.188	2.900	.005	.010	.055
	diet inventory (higher the score, higher cholesterol content)	.363	.114	.440	3.168	.002	.135	.591
	duration of exercise (hours/week)	-.511	.113	-.533	-4.519	.000	-.736	-.286
	socio-econmic status (index)	1.545E-02	.050	.055	.309	.758	-.084	.115
2	(Constant)	7.297	.620		11.763	.000	6.062	8.532
	age in year	3.281E-02	.011	.188	2.914	.005	.010	.055
	diet inventory (higher the score, higher cholesterol content)	.394	.052	.478	7.527	.000	.290	.498
	duration of exercise (hours/week)	-.540	.062	-.563	-8.702	.000	-.663	-.416

a. Dependent Variable: cholesterol in mmol/L

***P* values**

From the above 3 automatic procedures, we obtain the preliminary main effect model as:

Cholesterol = 7.297 – (0.540*exercise) + (0.394*diet) + (0.033*age)

Steps in Handling MLR

Step 1: Data exploration (Descriptive Statistics)

Step 2: Scatter plots and Simple Linear Regression

Step 3: Variable selection

⇨ Preliminary main-effect model

Step 4: Checking interaction & multicollinearity[a]

⇨ Preliminary final model

Step 5: Checking model assumptions & outliers[a]

⇨ Final model

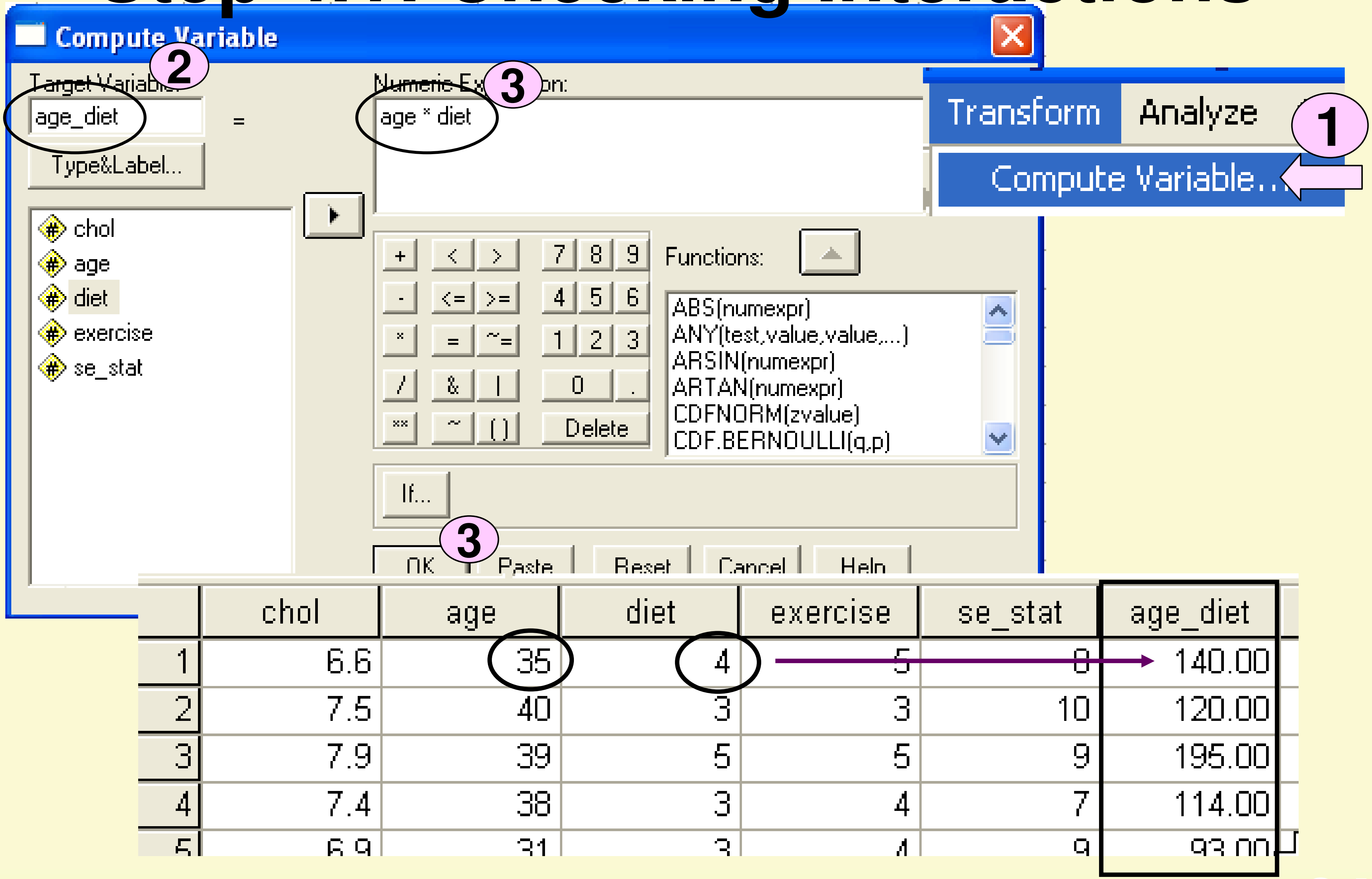
Step 6: Interpretation & data presentation

[a] need remedial measures if problems are detected

Step 4.1: Checking Interactions

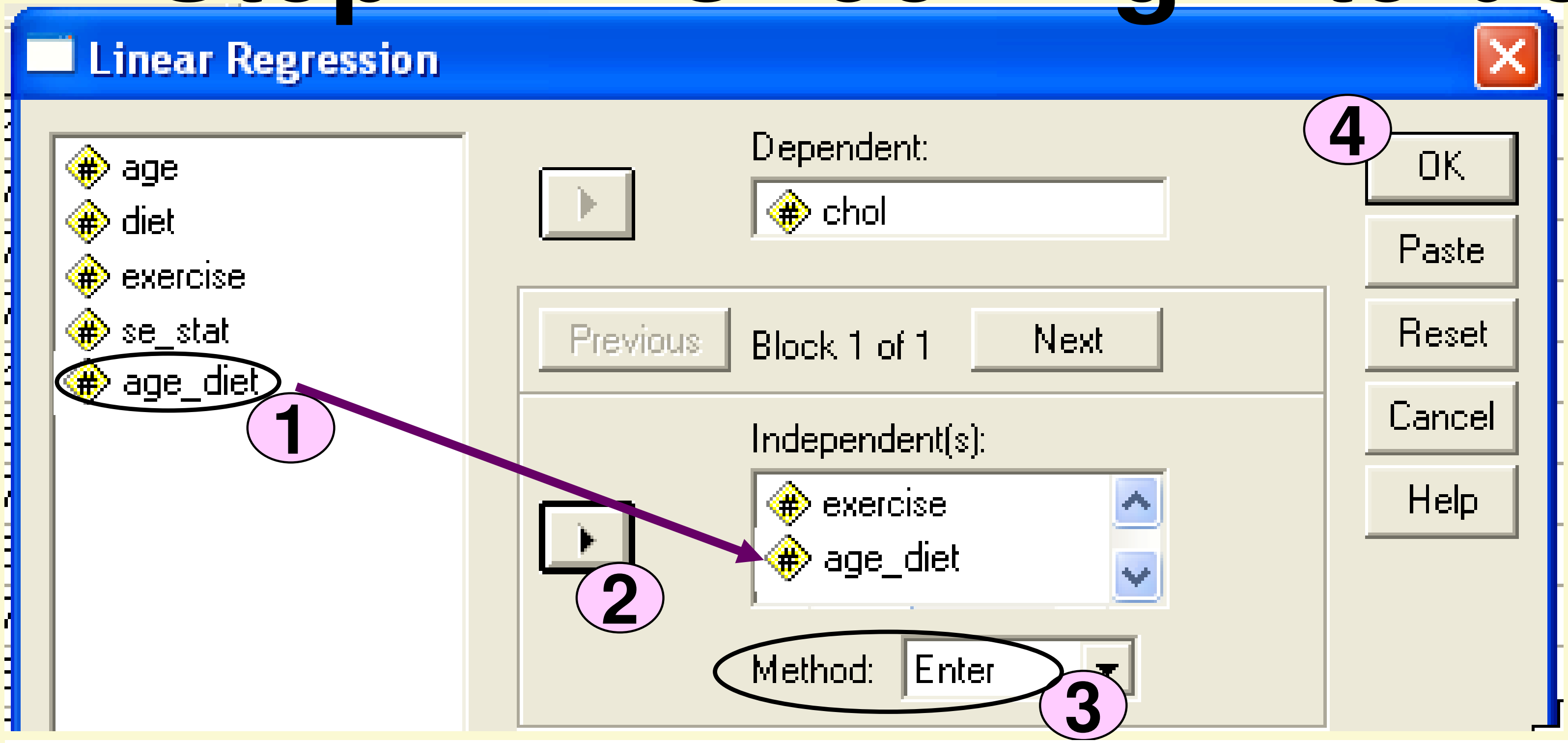
- **All possible 2-ways interactions (ex*diet; ex*age; diet*age) are checked.**
  - **Interaction terms are calculated (Transform⇨Compute).**
  - **Add into the model as additional independent variable.**
  - **Run the model using '*enter*'.**
  - **If an interaction term is significant (*P*<0.05), it means that there is an interaction between the 2 variables. And *therefore*, the appropriate model is the main effect variables plus the significant interaction term.**
  - **Check one interaction term at a time.**
- **In our example data, all 3 interaction terms are not significant. It means that <u>no interaction term</u> should be added.**

Step 4.1: Checking Interactions



	chol	age	diet	exercise	se_stat	age_diet
1	6.6	35	4	5	8	140.00
2	7.5	40	3	3	10	120.00
3	7.9	39	5	5	9	195.00
4	7.4	38	3	4	7	114.00
5	6.9	31	3	4	9	93.00

Step 4.1: Checking Interactions



Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95 Lo
1	(Constant)	6.812	2.102		3.240	.002	
	age in year	4.490E-02	.051	.257	.874	.385	
	diet inventory (higher the score, higher cholesterol content)	.495	.421	.600	1.176	.243	
	duration of exercise (hours/week)	-.539	.063	-.562	-8.610	.000	
	AGE_DIET	-2.53E-03	.010	-.144	-.241	.810	

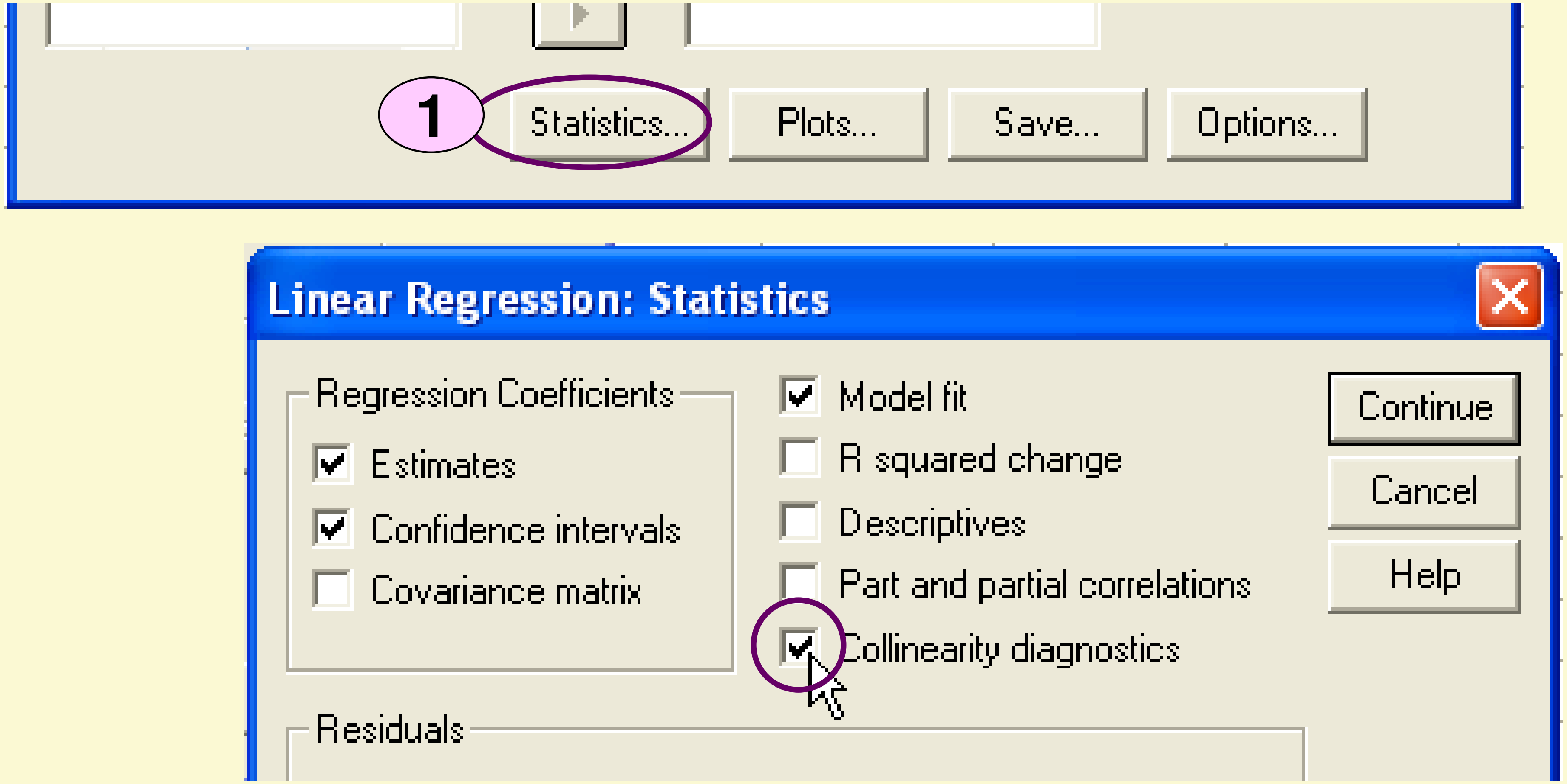
a. Dependent Variable: cholesterol in mmol/L

Step 4.2: Checking Multicollinearity (MC)

- If the independent variables are highly correlated, the regression model is said to be “statistically not stable”.
  - *P* values of the involved variables are considerably larger (than what it should be).
  - The width of 95% CI of the regression coefficients are larger.
  - Appropriate variables may be rejected wrongly.
  - Therefore, statistically, it is said that ‘the model is not stable’.
- We have to check the obtained model whether this kind of problem (MC) exists or not.

Step 4.2: Checking Multicollinearity (MC)

- **Just run the Preliminary main effect model by using 'enter', and click 'collinearity diagnostic' in 'statistics'.**



Step 4.2: Checking Multicollinearity (MC)

- **Just run the Preliminary main effect model by using ‘enter’, and click ‘collinearity diagnostic’ in ‘statistics’.**

Model		Unstandar Coefficie B	S	95% Confidence Interval for B Lower Bound	Upper Bound	Collinearity Statistics Tolerance	VIF
1	(Constant)	7.297	0	6.062	8.532		
	age in year	3.281E-02	05	.010	.055	.957	1.045
	diet inventory (higher the score, higher cholesterol content)	.394	0	.290	.498	.988	1.012
	duration of exercise (hours/week)	-.540	0	-.663	-.416	.950	1.053

a. Dependent Variable: cholesterol in mmol/L

Look at VIF (Variance-inflation factor). VIF measures the extent of multicollinearity problem. If VIF is more than 10, the problem needs remedial measures. Consult a statistician.

Steps in Handling MLR

Step 1: Data exploration (Descriptive Statistics)

Step 2: Scatter plots and Simple Linear Regression

Step 3: Variable selection

⇨ Preliminary main-effect model

Step 4: Checking interaction & multicollinearity[a]

⇨ Preliminary final model

Step 5: Checking model assumptions & outliers[a]

⇨ Final model

Step 6: Interpretation & data presentation

[a] need remedial measures if problems are detected

Step 5: Checking model assumptions

- Assumptions are …
  - Random sample*
  - Linearity → Overall linearity / Model fitness; Linearity of each indep. variable
  - Independence*
  - Normality
  - Equal variance

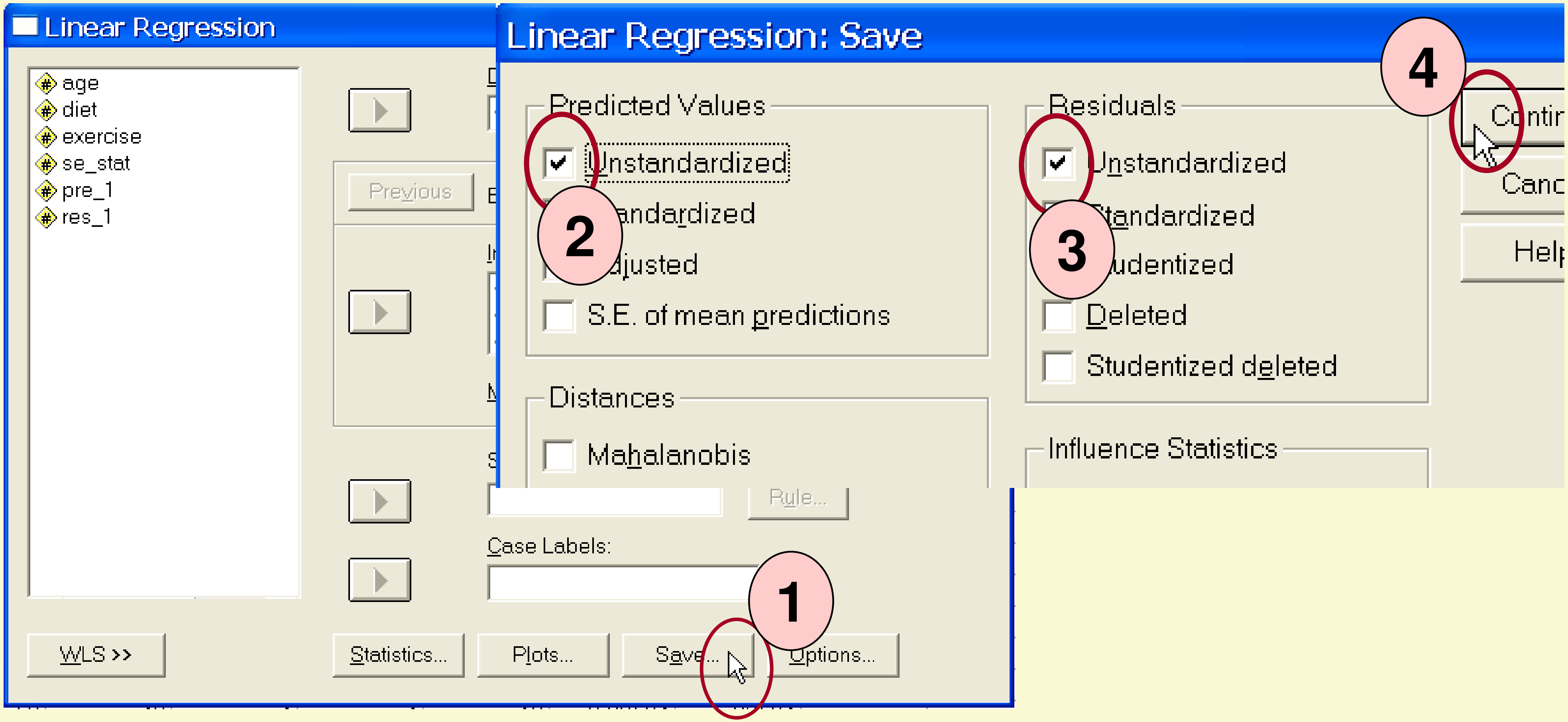
LINE (Linearity, Independence, Normality, Equal variance)

* It is related to the study design.

- All are performed by using residual plots.
- A residual means "observed value" minus "predicted value" of dependent variable.

Step 5: Checking model assumptions

Steps to calculate residuals …



Step 5: Checking model assumptions

	age	diet	exercise	chol	pre_1	res_1
1	35	4	5	6.6	7.32127	-.72127
2	40	3	3	7.5	8.17117	-.67117
3	39	5	5	7.9	7.84649	.05351
4	38	3	4	7.4	7.56563	-.16563
5	31	3	4	6.9	7.33597	-.43597
6	31	5	4	8.3	8.12397	.17603
7	38	6	5	7.6	8.20768	-.60768
8	48	4	3	8.9	8.82763	.07237
9	39	5	5	7.9	7.84649	.05351
10	38	7	5	8.6	8.60168	-.00168

Chol (*pred.*) = 7.297 – (0.540*exercise) + (0.394*diet) + (0.033*age)

Chol (*pred.*) = 7.297 – (0.540*5) + (0.394*4) + (0.033*35)

Chol (*pred.*) = 7.32

Residual = Chol (observed) – Chol (pred.) = 6.6 – 7.32 = – 0.72

Step 5: Checking model assumptions

Data | Statistical Model | Discrepancy

	age	diet	exercise	chol	pre_1	res_1
1	35	4	5	6.6	7.32127	-.72127
2	40	3	3	7.5	8.17117	-.67117
3	39	5	5	7.9	7.84649	.05351
4	38	3	4	7.4	7.56563	-.16563
5	31	3	4	6.9	7.33597	-.43597
6	31	5	4	8.3	8.12397	.17603
7	38	6	5	7.6	8.20768	-.60768
8	48	4	3	8.9	8.82763	.07237
9	39	5	5	7.9	7.84649	.05351
10	38	7	5	8.6	8.60168	-.00168

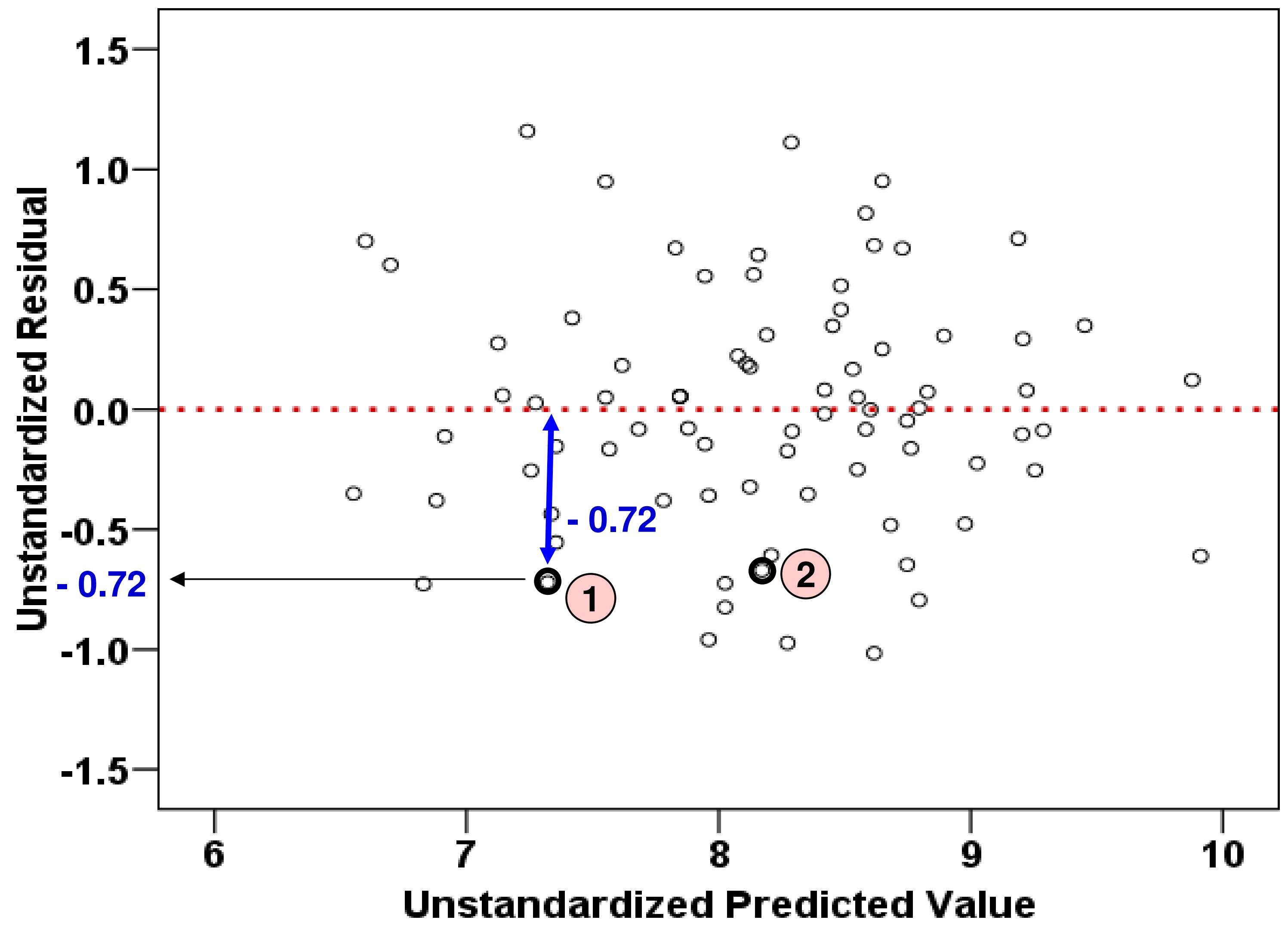
Chol (*pred.*) = 7.297 – (0.540*exercise) + (0.394*diet) + (0.033*age)

Chol (*pred.*) = 7.297 – (0.540*5) + (0.394*4) + (0.033*35)

Chol (*pred.*) = 7.32

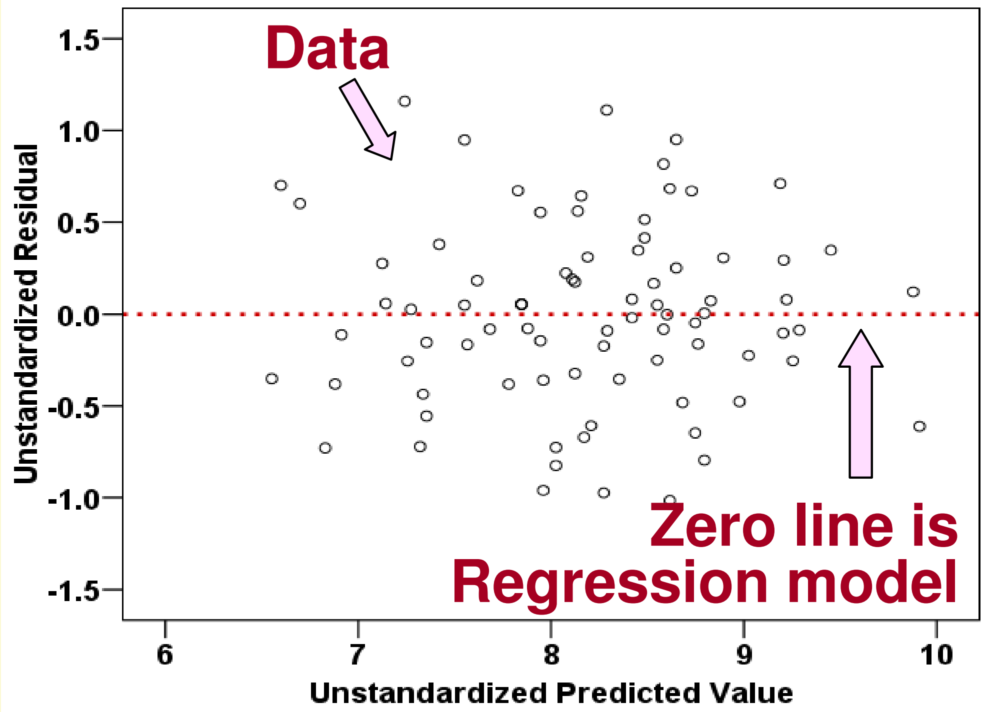
Residual = Chol (observed) – Chol (pred.) = 6.6 – 7.32 = – 0.72

Step 5: Checking model assumptions



	age	diet	exercise	chol	pre_1	res_1
1	35	4	5	6.6	7.32127	-.72127
2	40	3	3	7.5	8.17117	-.67117

Step 5: Checking model assumptions



Simple Linear Regression

Multiple Linear Regression

Step 5: Checking model assumptions

- Assumptions are …
  - Random sample*
  - Linearity → Overall linearity / Model fitness; Linearity of each indep. variable
  - Independence*
  - Normality
  - Equal variance

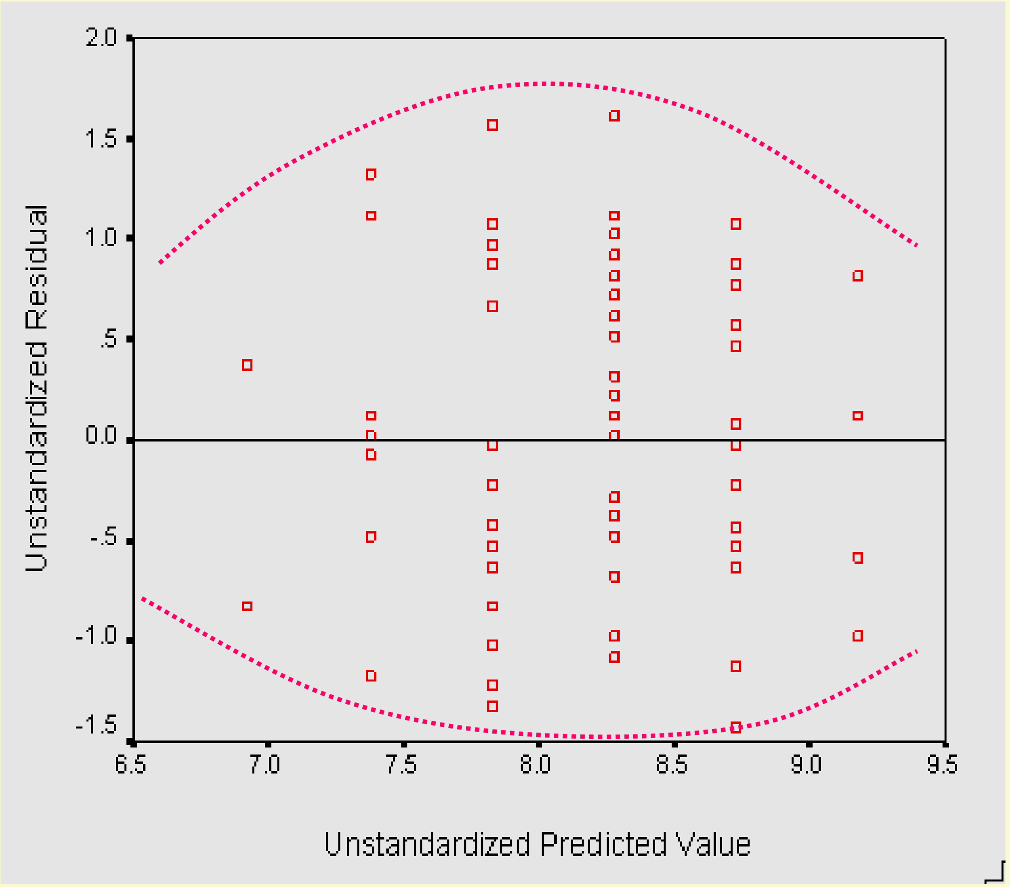
LINE

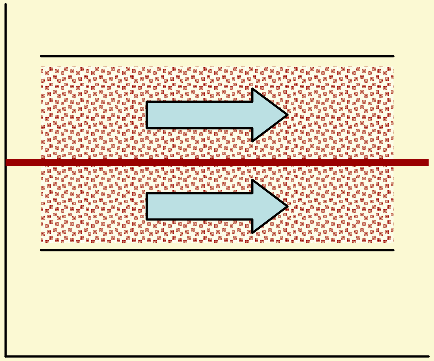
* It is related to the study design.

	3 types of residual plot	Assumptions
1.	Scatter plot: Residuals vs Predicted	Linearity – overall fitness Equal variance of residuals
2.	Histogram of residuals	Normality of residuals
3.	Scatter plot: Residuals vs each indep. var. (numerical)	Linearity of each indep. Var. numerical

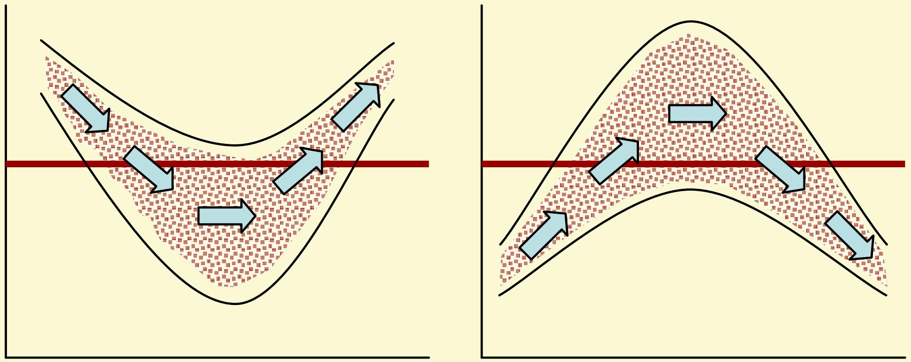
Step 5: Checking model assumptions

OVERALL LINEARITY



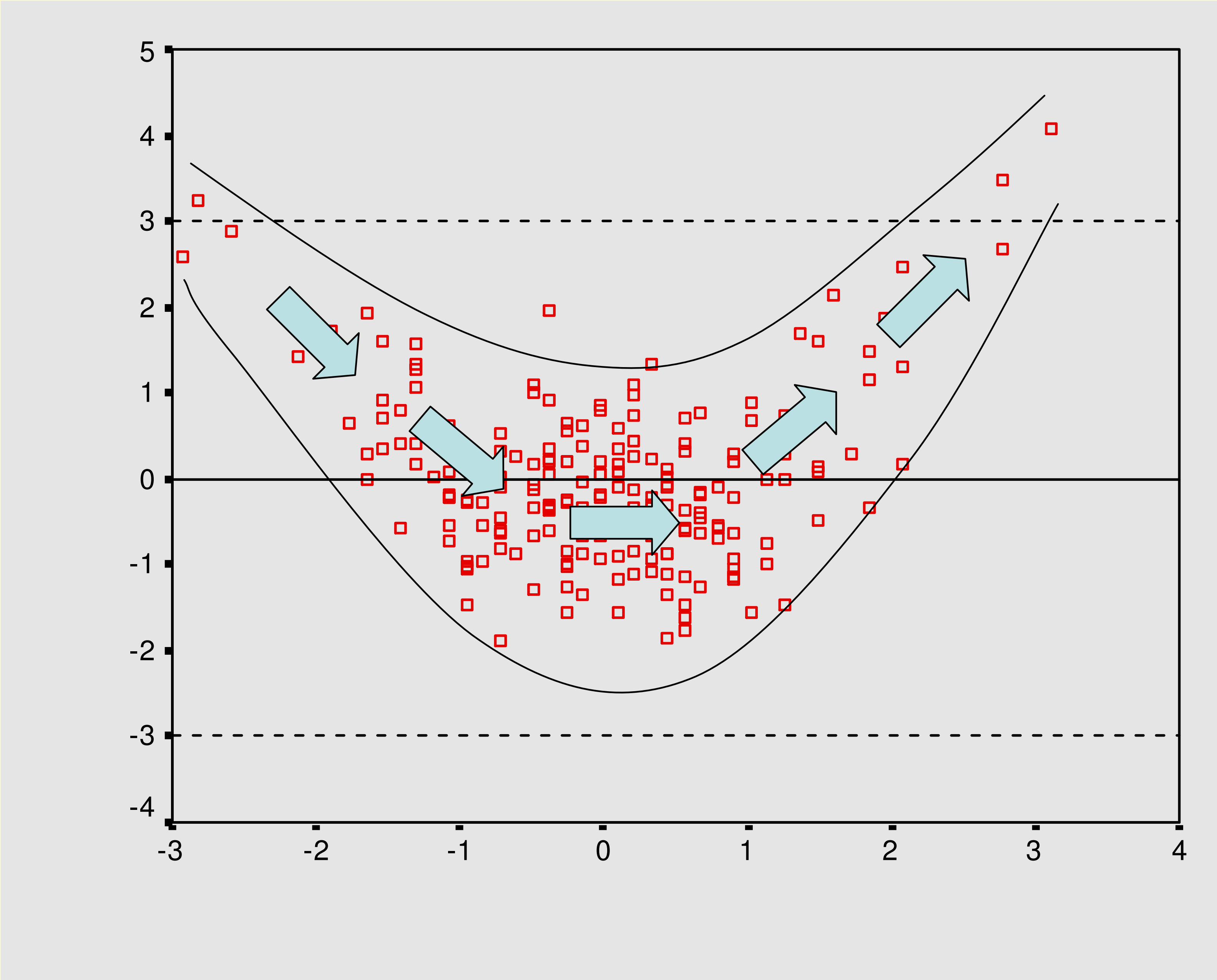


Linearity assumption is met (linear model fits well).

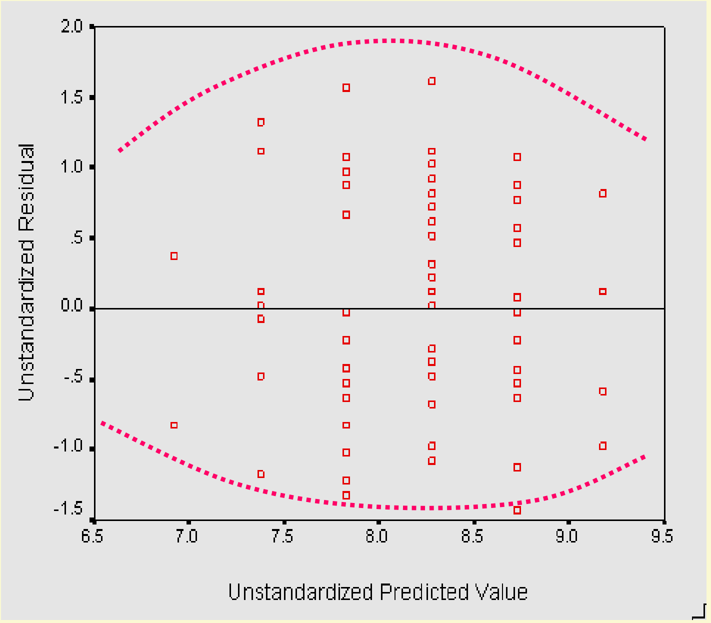


Linear assumption is not met (linear model doesn't fit well).

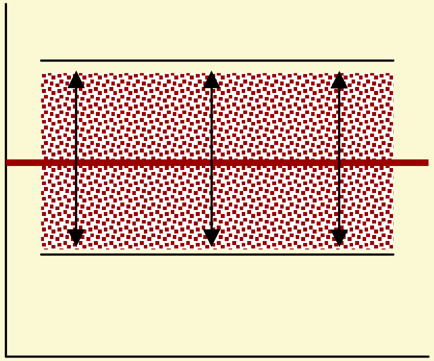
An example of non-linear relationship



Step 5: Checking model assumptions

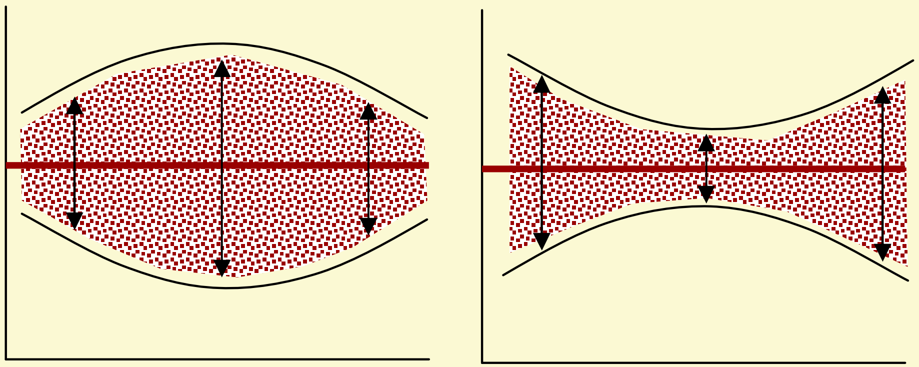


EQUAL VARIANCE

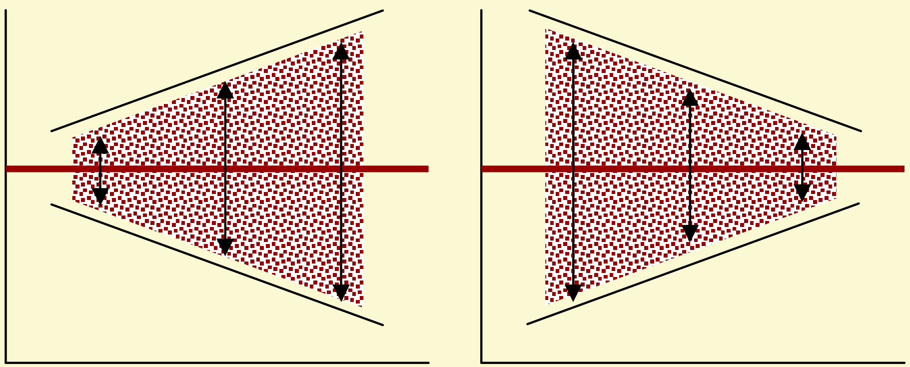


Equal-variance assumption is met.

"Constantly increasing or decreasing"



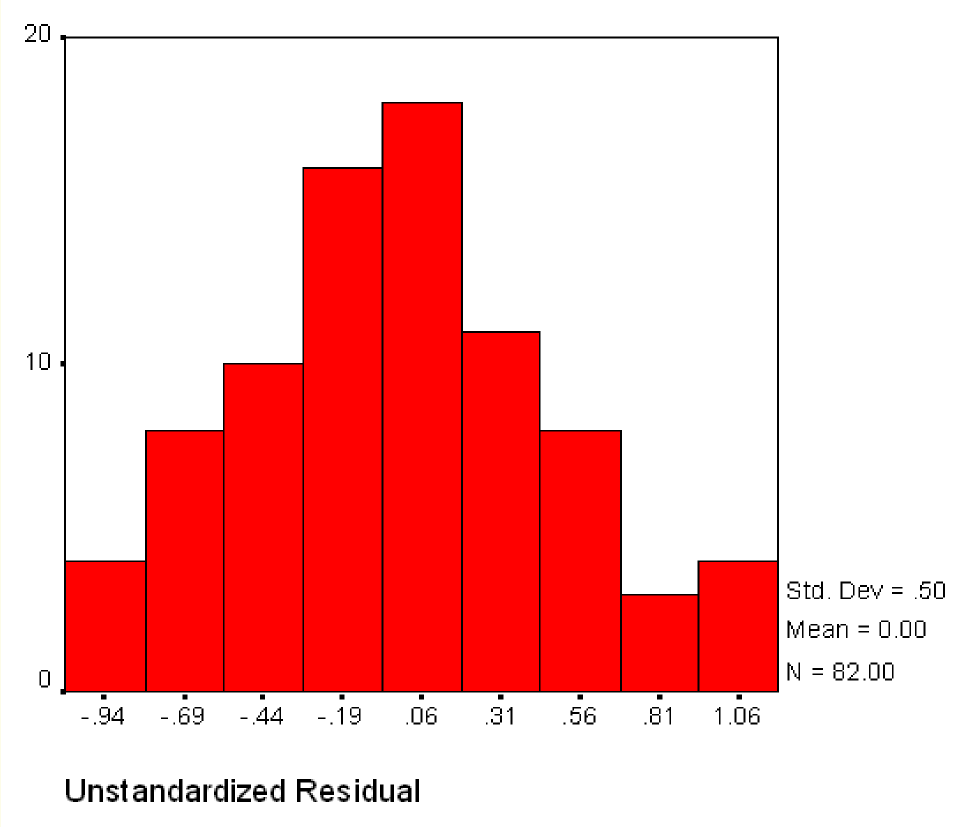
Equal-variance assumption is met.

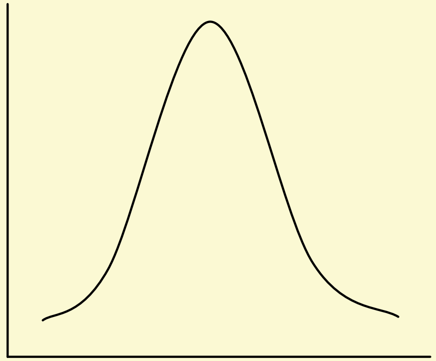


Equal-variance assumption is NOT met.

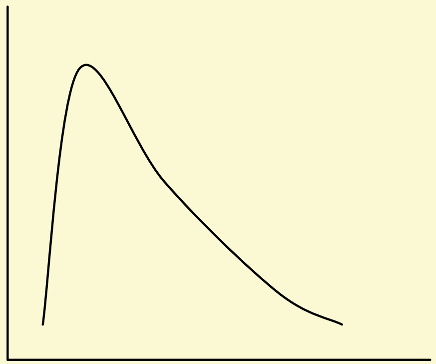
Step 5: Checking model assumptions

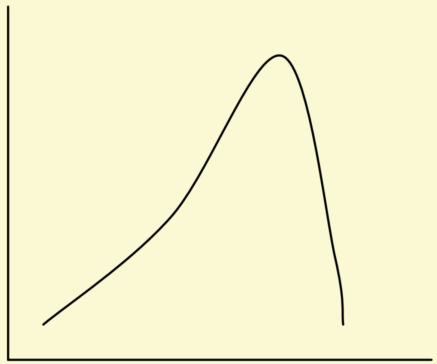
NORMALITY





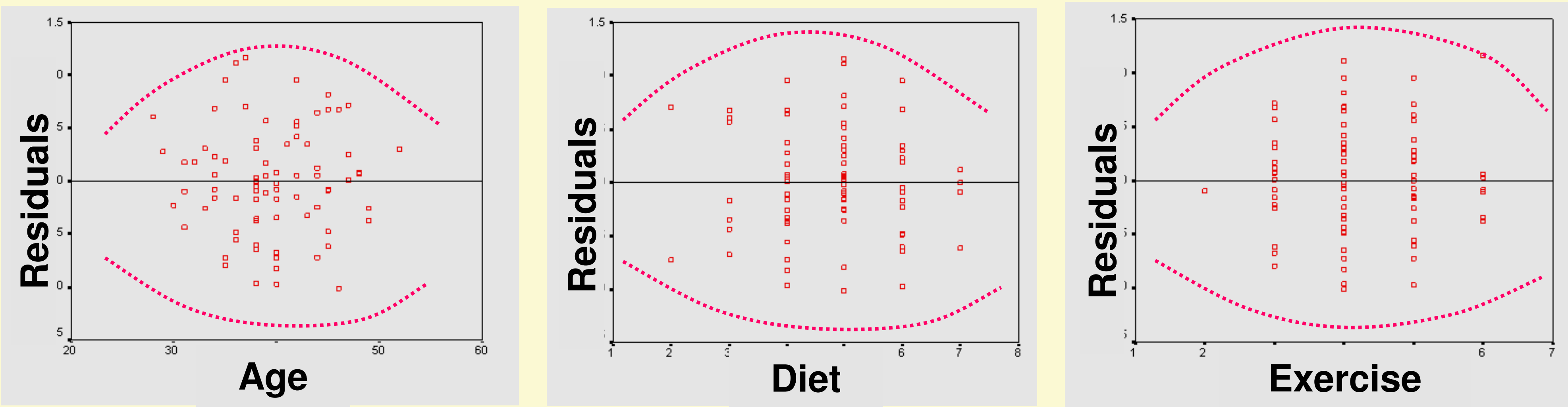
Normality assumption is met.





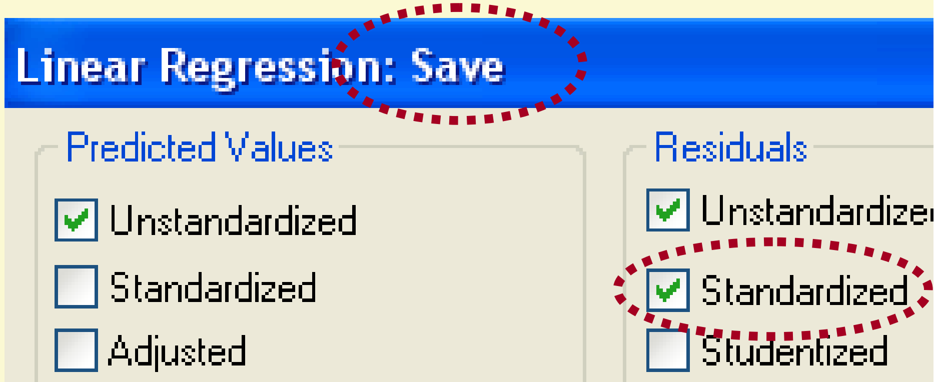
Normality assumption is not met.

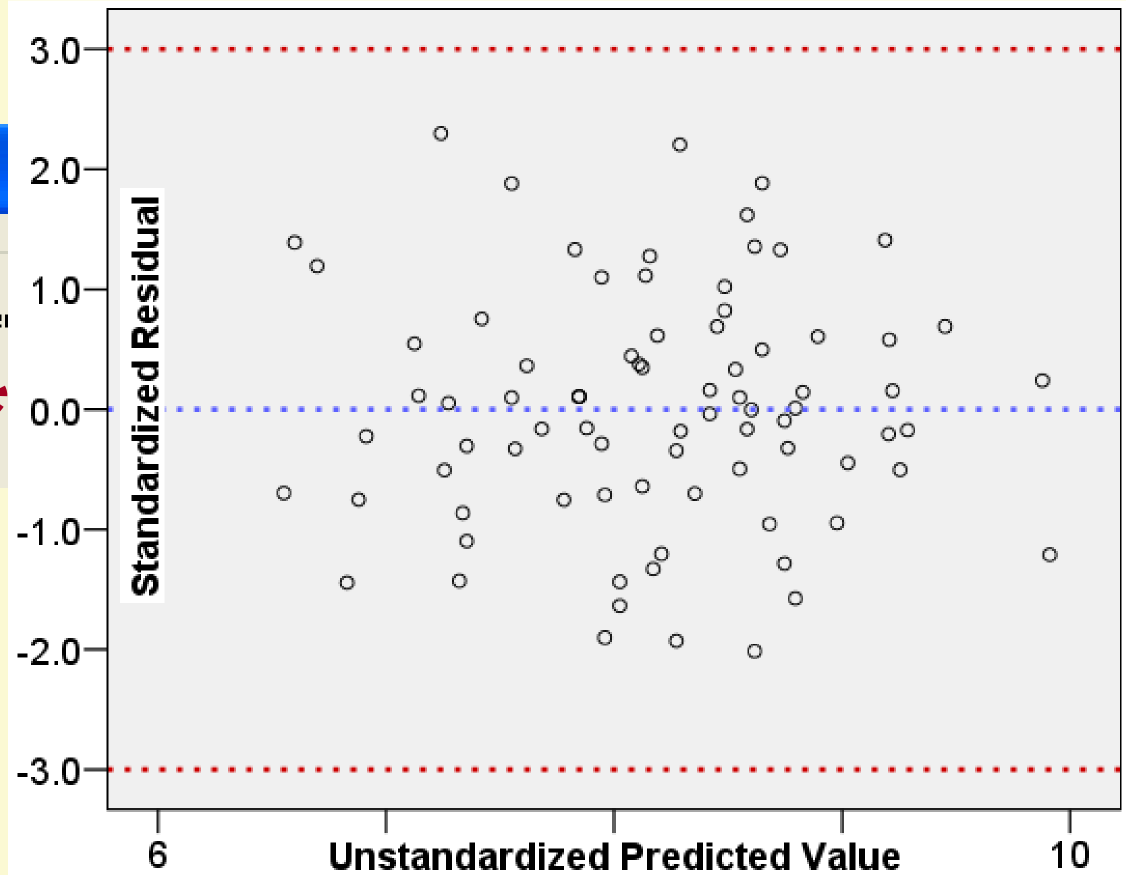
Checking linearity of each numerical independent variables



- **If there is no relationship between residuals and a numerical independent variable, the relationship of the independent variable with the outcome is linear.**
- **In above example, all are considered linear relationship.**
- **If not linear, we may need to transform data (*see statistician*).**

Checking Outliers





- If data points are beyond +3 and -3 of standardized residuals, they are considered "outliers".
- Check for 'data entry error' and 'eligibility' as study subjects. If no entry error and are an eligible cases, consult a statistician to handle the outliers.

Steps in Handling MLR

Step 1: Data exploration (Descriptive Statistics)

Step 2: Scatter plots and Simple Linear Regression

Step 3: Variable selection

⇨ Preliminary main-effect model

Step 4: Checking interaction & multicollinearity[a]

⇨ Preliminary final model

Step 5: Checking model assumptions & outliers[a]

⇨ **Final model**

Step 6: Interpretation & data presentation

[a] need remedial measures if problems are detected

Step 6: Presentation/Interpretation

Table 4: Factors associated with blood cholesterol level (mmol/L) among the study population (*n*=82)

Variables	SLR[a]			MLR[b]			
	b[c]	(95% CI)	*P* value	*Adj.b*[d]	(95%CI)	*t*-stat.	*P* value
Age (year)	0.06	(0.02, 0.09)	0.002	0.03	(0.01, 0.06)	2.91	0.005
Duration of exercise (hrs/wk)	-0.62	(-0.79, -0.46)	<0.001	-0.54	(-0.66, -0.42)	- 8.70	<0.001
Diet inventory score	0.45	(0.30, 0.61)	<0.001	0.39	(0.29, 0.50)	7.53	<0.001
Socio-economic index	0.21	(0.17, 0.25)	<0.001	-	-	-	-

[a] Simple linear regression
[b] Multiple linear regression (R^2=0.69; The model reasonably fits well; Model assumptions are met; There is no interaction between independent variables, and no multicollinearity problem)
[c] Crude regression coefficient
[d] Adjusted regression coefficient

- **For prediction study, it is essential to report the final model (equation).**

Chol (*pred.*) = 7.30 + (0.03age*) – (0.54**exercise*) + (0.39**diet*)**

Step 6: Presentation/Interpretation

Table 4: Factors associated with blood cholesterol level (mmol/L) among the study population (*n*=82)

Variables	SLR[a]			MLR[b]			
	b[c]	(95% CI)	*P* value	*Adj.b*[d]	(95%CI)	*t*-stat.	*P* value
Age (year)	0.06	(0.02, 0.09)	0.002	0.03	(0.01, 0.06)	2.91	0.005
Duration of exercise (hrs/wk)	-0.62	(-0.79, -0.46)	<0.001	-0.54	(-0.66, -0.42)	- 8.70	<0.001
Diet inventory score	0.45	(0.30, 0.61)	<0.001	0.39	(0.29, 0.50)	7.53	<0.001
Socio-economic index	0.21	(0.17, 0.25)	<0.001	-	-	-	-

- **There is a significant linear relationship between age and cholesterol level (*P*=0.005). Those with 10 years older have cholesterol level higher for 0.3 mmol/L (95% CI: 0.1, 0.6 mmol/L).**
- **There is a significant linear relationship between duration of exercise and cholesterol level (*P*<0.001). Those having 1 hr/wk less exercise have cholesterol level higher for 0.54 mmol/L (95% CI: 0.66, 0.42 mmol/L).**

Step 6: Presentation/Interpretation

Table 4: Factors associated with blood cholesterol level (mmol/L) among the study population (*n*=82)

Variables	SLR[a]			MLR[b]			
	b[c]	(95% CI)	*P* value	*Adj.b*[d]	(95%CI)	*t*-stat.	*P* value
Age (year)	0.06	(0.02, 0.09)	0.002	0.03	(0.01, 0.06)	2.91	0.005
Duration of exercise (hrs/wk)	-0.62	(-0.79, -0.46)	<0.001	-0.54	(-0.66, -0.42)	- 8.70	<0.001
Diet inventory score	0.45	(0.30, 0.61)	<0.001	0.39	(0.29, 0.50)	7.53	<0.001
Socio-economic index	0.21	(0.17, 0.25)	<0.001	-	-	-	-

- **There is a significant linear relationship between diet inventory index and cholesterol level (*P*<0.001). Those with 1 unit more in the index, have cholesterol level higher for 0.39 mmol/L (95% CI: 0.29, 0.50 mmol/L).**
- **With the 3 significant variables, the model explains 69% of variation of the blood cholesterol level in the study sample. (R^2=0.69)**

SUMMARY

Step 1: Data exploration (Descriptive Statistics)
Step 2: Scatter plots and Simple Linear Regression
} **Exploring**

Step 3: Variable selection
⇨ **Preliminary main-effect model**
Step 4: Checking interaction & multicollinearity[a]
⇨ **Preliminary final model**
} **Modeling**

Step 5: Checking model assumptions & outliers[a]
⇨ **Final model**
Step 6: Interpretation & data presentation
} **Checking assumptions and interpretation**

[a] need remedial measures if problems are detected.

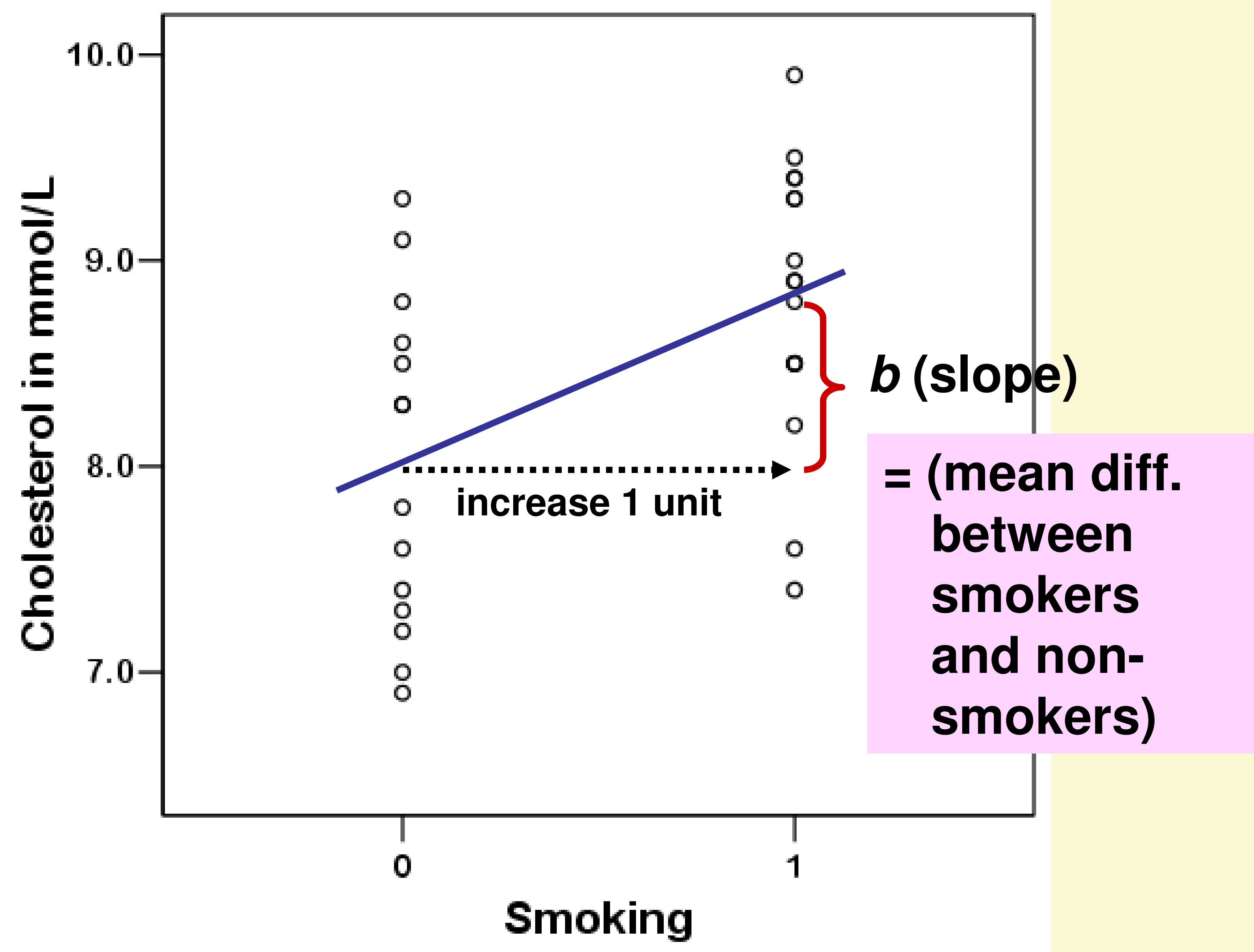
Categorical Independent Var.

Cautions:
It should be coded (0, 1) for dichotomous variable.

Example 1: sex (male=1, female=0)
It means we are comparing male against female (female as reference)

Example 2: smoking (smokers=1, non-smoker=0)
It means we are comparing smokers against non-smoker (non-smoker as reference)

Say, outcome is cholesterol, smoking as independent var., and we got *b*=2.0. It means smokers will have cholesterol level higher than non-smokers for 2.0 mmol/L.


10.0
9.0
8.0
7.0
Cholesterol in mmol/L
increase 1 unit
0
1
Smoking
b (slope)
= (mean diff. between smokers and non-smokers)

Categorical Independent Var.

Cautions:

If you have more than 2 categories in categorical variable, we have to create <u>Dummy Variables</u>.

Example: Education level (no education=1; primary school level=2; secondary level=3)

Then, we need to create 2 dummy variables: (e.g. <u>edu2</u> & <u>edu3</u>)

	edu2	edu3
No edu. →	0	0
Primary edu. →	1	0
Secondary edu. →	0	1

Here, reference is 'no education',

edu2 is comparing 'primary' against 'no edu', and

edu3 is comparing 'secondary' against 'no edu'.

Categorical Independent Var.

Example 2: Education level (no education=1; primary=2; secondary=3; tertiary=4)

Then, we need to create 3 dummy variables: (e.g. edu2 & edu3 & edu4)

	edu2	edu3	edu4
No edu. →	0	0	0
Primary edu. →	1	0	0
Secondary edu. →	0	1	0
Tertiary edu. →	0	0	1

Categorical Independent Var.

Cautions:

If you have more than 2 categories in categorical variable, we have to create <u>Dummy Variables</u>.

Example: Agegp: Age (<35)=1; Age (35-44)=2; Age (>=45)=3

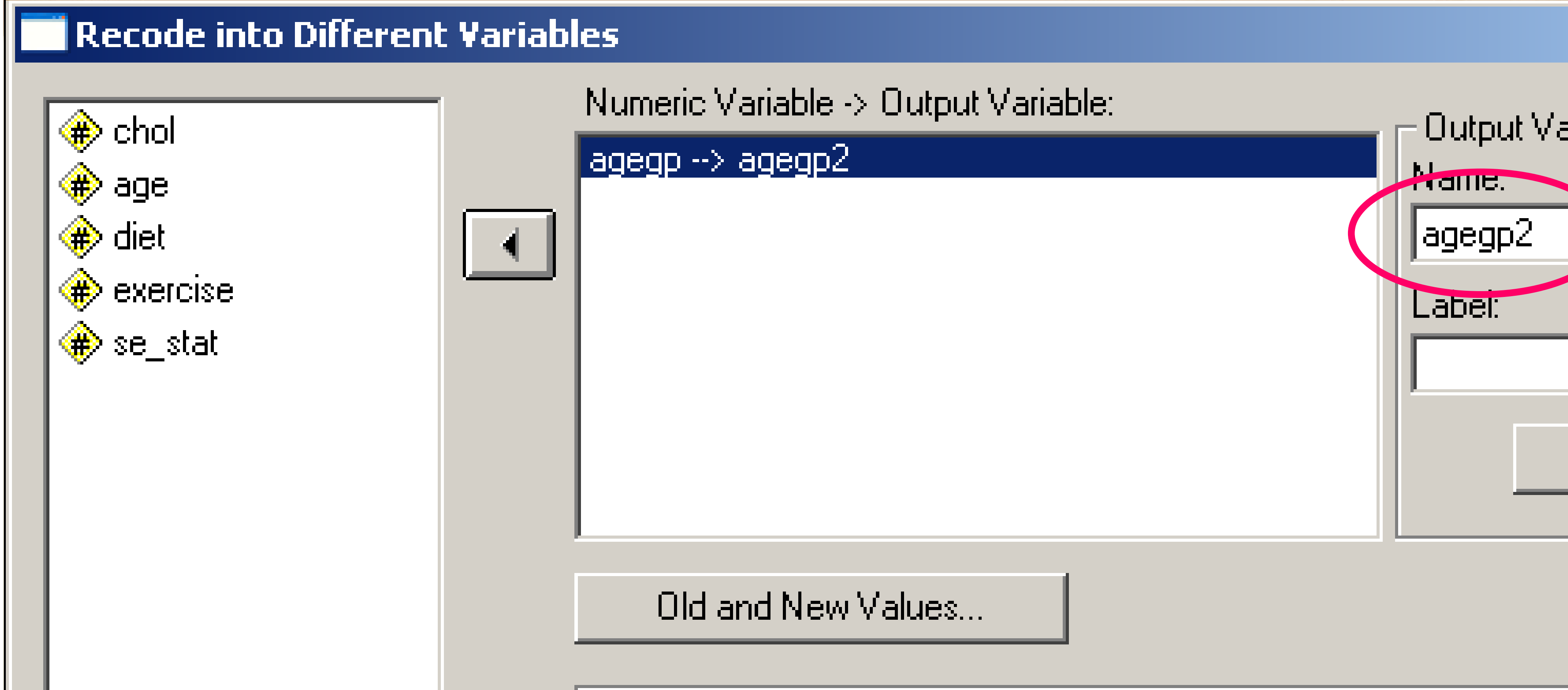
Then, we need to create 2 dummy variables: (e.g. <u>agegp2</u> & <u>agegp3</u>)

agegp		agegp2	agegp3
<35 (1) 'yg'	→	0	0
35-44 (2) 'older'	→	1	0
>=45 (3) 'eldest'	→	0	1

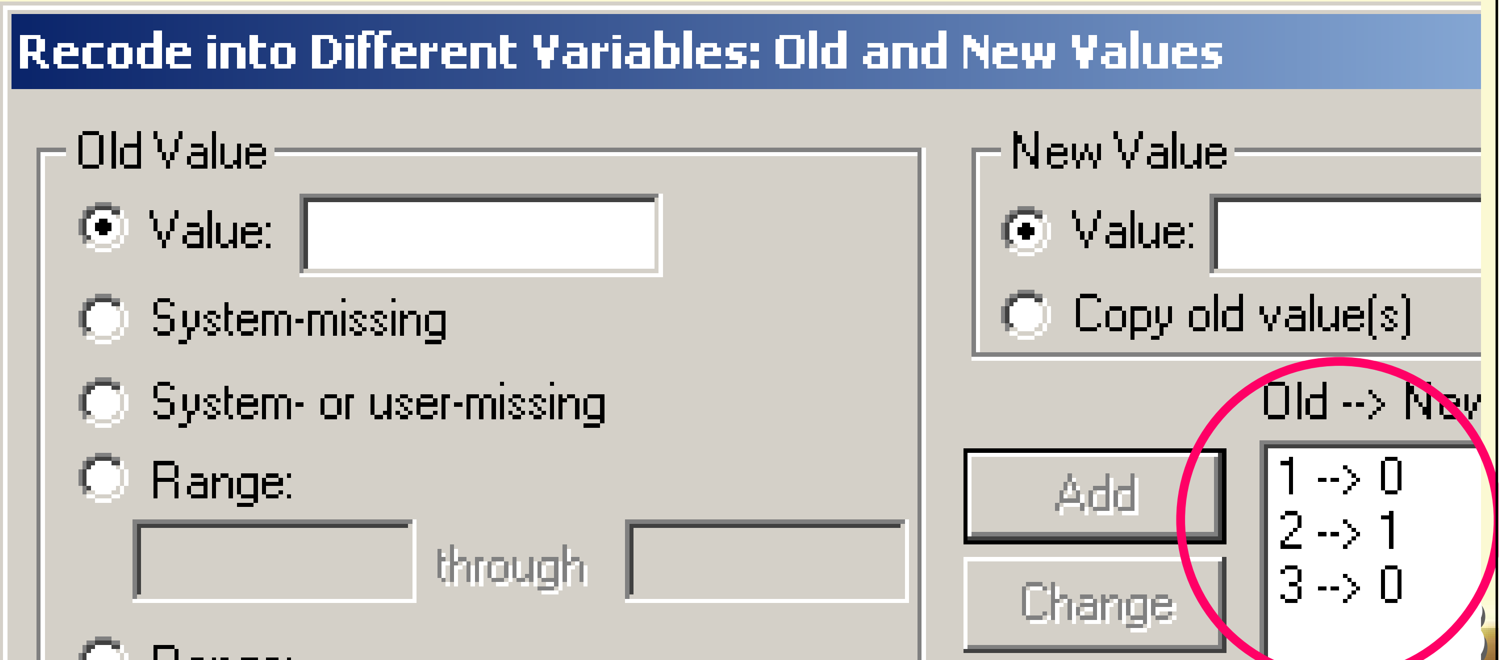
Here, reference is 'young',

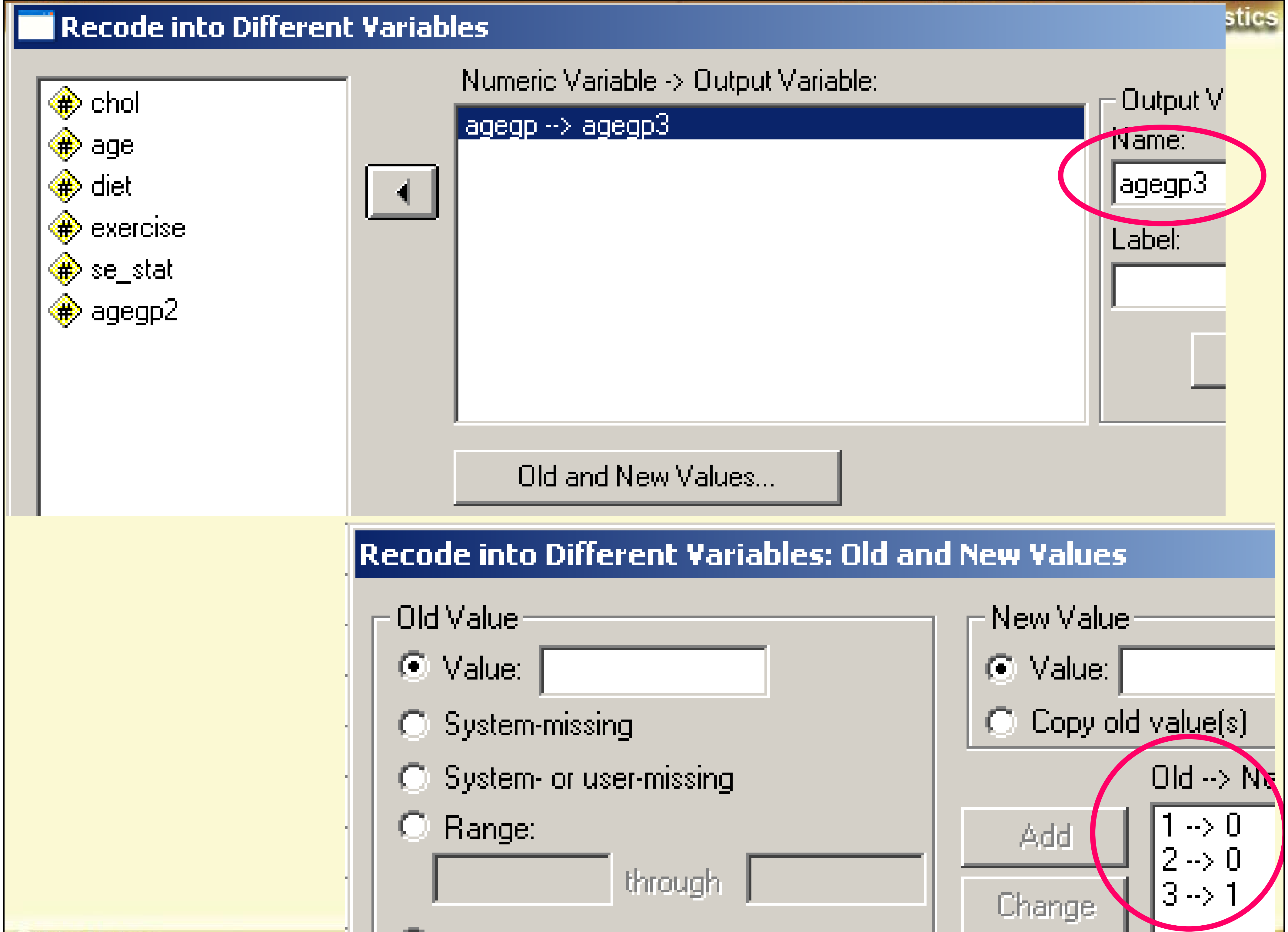
agegp2 is comparing 'older' against 'young', and

agegp3 is comparing 'eldest' against 'young'.

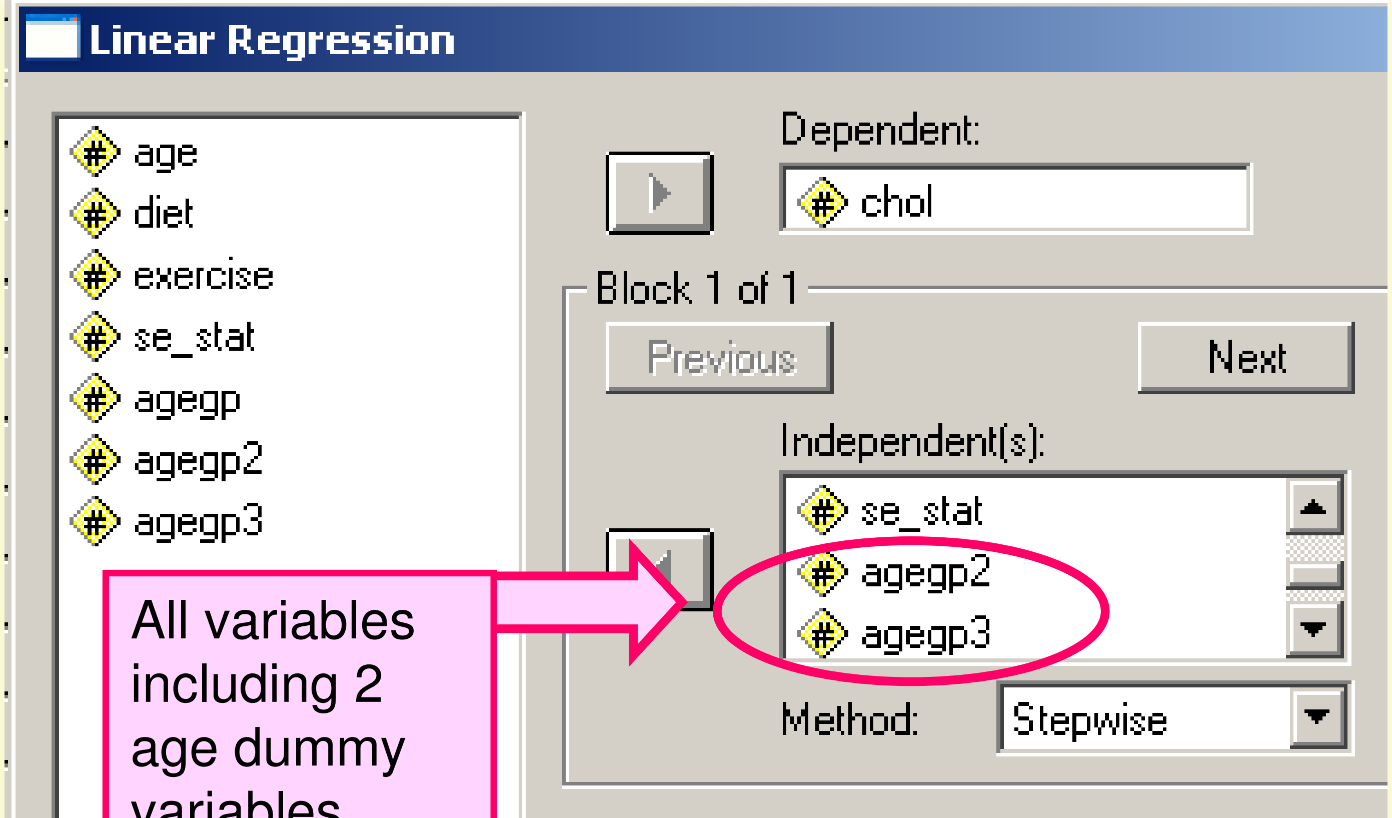


‘Recode’ into different variables




Recode into Different Variables
chol
age
diet
exercise
se_stat
agegp2
Numeric Variable -> Output Variable:
agegp --> agegp3
Output V
Name:
agegp3
Label:
Old and New Values...
Recode into Different Variables: Old and New Values
Old Value
Value:
System-missing
System- or user-missing
Range:
through
New Value
Value:
Copy old value(s)
Add
Change
1 --> 0
2 --> 0
3 --> 1

Variable selection procedure

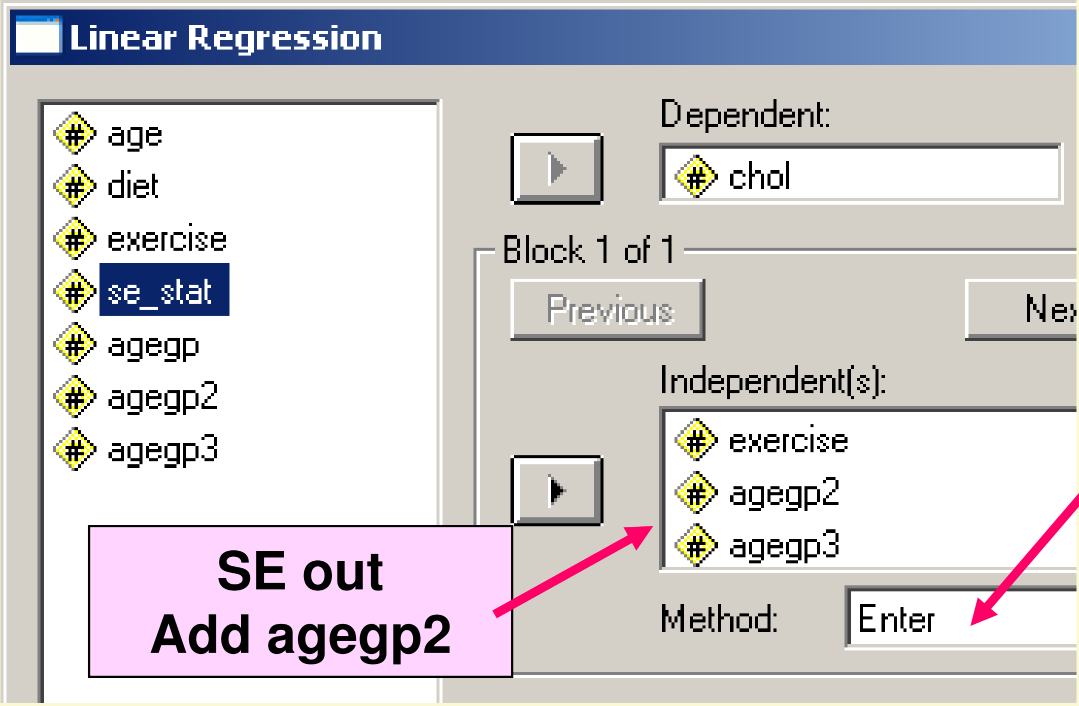


All variables including 2 age dummy variables

P value

5	(Constant)	8.529	.397		21.473	.000	7.738	9.3
	exercise	-.536	.064	-.559	-8.368	.000	-.664	-.4
	diet	.390	.053	.473	7.310	.000	.284	.4
	agegp3	.351	.149	.157	2.354	.021	.054	.6

SE is out, and only agegp3 is selected. However, agegp3 is part of age variable, and both dummy variables must be in the model (to complete as the age variable).



We have to force agegp2 to complete as the age-group variable.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	8.445	.415		20.373	.000
	diet	.391	.054	.474	7.302	.000
	exercise	-.539	.064	-.562	-8.369	.000
	agegp2	.114	.155	.062	.737	.464
	agegp3	.439	.192	.197	2.291	.025

a. Dependent Variable: chol

How to interpret '*b*' of categorical variable?

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	8.445	.415		20.373	.000
	diet	.391	.054	.474	7.302	.000
	exercise	-.539	.064	-.562	-8.369	.000
	agegp2	.114	.155	.062	.737	.464
	agegp3	.439	.192	.197	2.291	.025

a. Dependent Variable: chol

- **There is no significant difference in cholesterol level between older age-group (35-44) and young group (<35) (*P*=0.464).**
- **However, the eldest group (>=45) have significantly higher cholesterol level than the young group (<35) (*P*=0.025).**
- **The eldest group (>=45) have 0.44 mmol/L higher cholesterol level than the young group (<35) (95% CI: 0.06, 0.82 mmol/L).**

Questions & Answers