Biostatistics 201: Linear Regression Analysis

Y H Chan

In the 100 series[1-4] the common **univariate** techniques (summarized in Table I) available for data analyses were discussed. These techniques do not allow us to take into account the effect of other covariates/ confounders (except for partial correlation[4]) in an analysis. In such situations, a **Regression Model** would be required.

Table I. Univariate Statistical techniques.

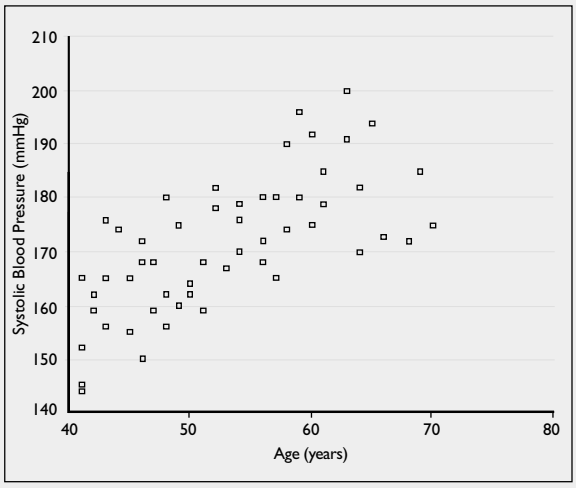
Quantitative Data	
Parametric tests	Non-Parametric tests
1 Sample T-test	Sign Test Wilcoxon Signed Rank test
Paired Sample t-test	
2 Sample T-test	Mann Whitney U test Wilcoxon Rank Sum test
One Way ANOVA	Kruskal Wallis test
Qualitative Data	
For independent Samples: Chi Square / Fisher's Exact test	
For Matched Case-Control Samples : McNemar Test	
Bivariate Correlation (Quantitative data)	
Normality assumptions satisfied	Pearson's Correlation
Normality assumptions not satisfied or Ordinal Qualitative data	Spearman's Correlation
Agreement Analysis	
Quantitative data	Bland Altman Plots
Qualitative data	Kappa Estimates

Reasons why we want a Regression Model

1. Descriptive - form the strength of the association between outcome and factors of interest
2. Adjustment - for covariates/confounders
3. Predictors - to determine important risk factors affecting the outcome
4. Prediction - to quantify new cases

In this article, we shall discuss the Regression modeling for a quantitative response outcome. For example, data (n = 55) on the age and the systolic BP were collected and we want to set-up a **Linear Regression Model** to predict BP with age. Here we could, after checking the normality assumptions for both variables, do a bivariate correlation (Pearson's correlation = 0.696, p<0.001) and a graphical scatter plot would be helpful (see Fig. 1).

Fig. 1 Scatter plot of Systolic BP versus Age.



There's a moderately strong correlation between age and systolic BP but how could we 'quantify' this strength.

SIMPLE LINEAR REGRESSION ANALYSIS (HAVING ONLY ONE PREDICTOR)

A simple linear regression model to relate BP with age will be

BP = regression estimate (b) * age + constant (a) + error term (å)

The regression estimate (b) and the constant (a) will be derived from the data (using the method of least-squares[5]) and the error term is to factor in the situation that two persons with the same age need not have the same BP.

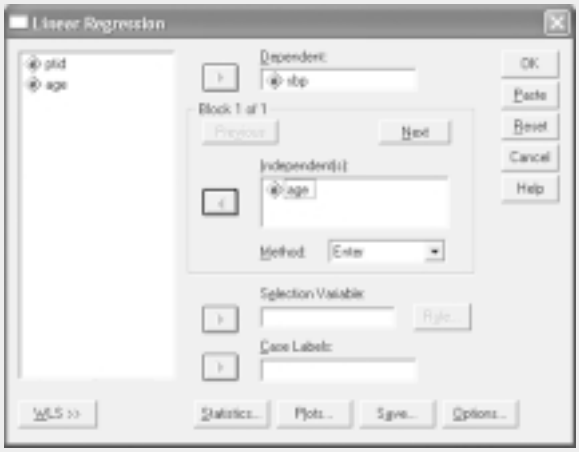
In SPSS (11.5), to perform a linear regression, go to **Analyse, Regression, Linear** to get template I.

Clinical Trials and Epidemiology Research Unit
226 Outram Road
Blk B #02-02
Singapore 169039

Y H Chan, PhD
Head of Biostatistics

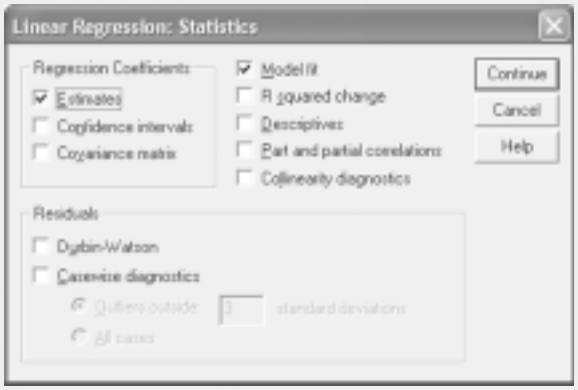
Correspondence to:
Dr Y H Chan
Tel: (65) 6325 7070
Fax: (65) 6324 2700
Email: chanyh@cteru.com.sg

Template I. Linear Regression Analysis.



Put sbp (systolic BP) as the Dependent and age as the Independent; click on the Statistics button to get template II.

Template II



Tick on the Confidence intervals box, continue and click OK in template I. Tables II a – d show the SPSS Simple Linear Regression outputs between Systolic BP and age.

Table IIa

Variables entered/removed[b]

Model	Variables Entered	Variables Removed	Method
1	Age (years)[a]	.	Enter

[a] All requested variables entered.

[b] Dependent variable: Systolic blood pressure (mmHg).

This table indicates the dependent and independent variables. The method of including the independent variable is Enter (see Model selection later)

Table IIb

Model summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.696[a]	.485	.475	9.10072

[a] Predictors: (Constant), Age (years)

Here the Pearson's correlation between SBP and age is given (r = 0.696). R square = 0.485 which implies that only 48.5% of the systolic BP is explained by the age of a person. We shall ignore the explanation for the adjusted R Square for the time being (see Multiple Linear Regression later).

Table IIc.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4128.118	1	4128.118	49.843	.000[a]
	Residual	4389.628	53	82.823		
	Total	8517.745	54			

[a] Predictors: (Constant), Age (years).

[b] Dependent variable: Systolic blood pressure (mmHg).

The ANOVA table shows the 'usefulness' of the linear regression model – we want the p-value to be <0.05.

Table IId

Coefficients[a]

	Unstandardised Coefficients		Standardised Coefficients			95% Confidence Interval for B	
Model	B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1 (Constant)	115.706	7.999		14.465	.000	99.662	131.749
Age (years)	1.051	.149	.696	7.060	.000	.752	1.350

[a] Dependent variable: Systolic blood pressure (mmHg).

Table IId provides the quantification of the relationship between age and systolic BP. With every increase of one year in age, the systolic BP (on the average) increases by 1.051 (95% CI 0.752 to 1.350) units, p<0.001. The Constant here has no 'practical' meaning as it gives the value of the systolic BP when age = 0. Sometimes we may want to make age 50 as reference. To do this, compute a new variable (age50 = age - 50). The constant in Table IIe gives the average systolic BP for a 50-year-old person : 168.3 (95% CI 165.6 to 170.9). Observe that the quantification of the relationship between age and systolic BP (b = 1.051) does not change with the 'new' model.

Table IIe. Age-centered at 50 years old.

Coefficients[a]

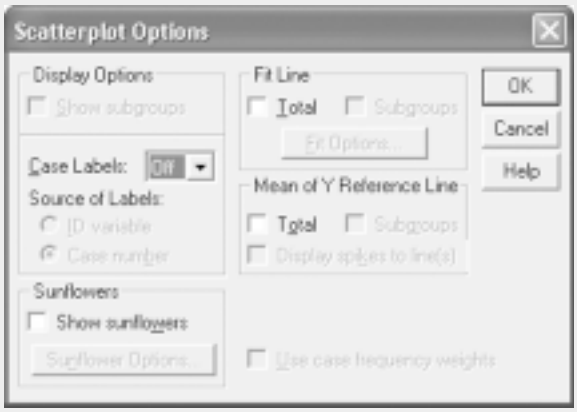
	Unstandardised Coefficients		Standardised Coefficients			95% Confidence Interval for B	
Model	B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1 (Constant)	168.260	1.311		128.385	.000	165.632	170.889
reference age = 50	1.051	.149	.696	7.060	.000	.752	1.350

[a] Dependent variable: Systolic blood pressure (mmHg).

For a single independent variable, the Standardised Coefficient (Beta) is the Pearson's correlation value (we shall discuss the use of Beta later in Multiple Regression).

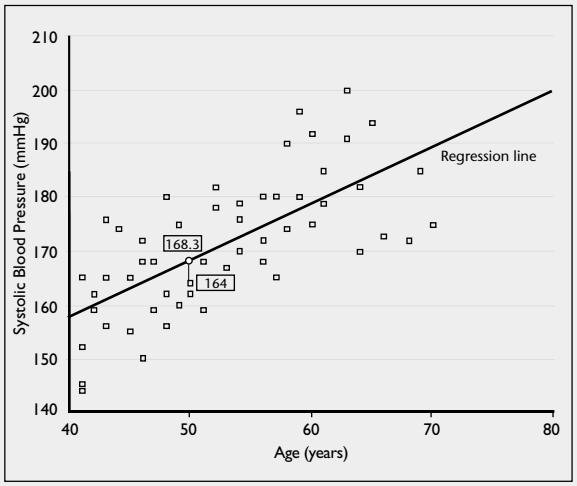
To include a regression line in the scatter plot, double-click on the plot to get into the Chart editor. Go to Chart, Options to get template III :

Template III



Tick the Fit Line Total box and Figure II will be obtained.

Fig. II Scatter plot with Regression line.



The equation of the Regression line is SBP = 115.706 + 1.105 * Age (see Table IId). We can use this descriptive relationship to predict the systolic BP for any age, between 40 to 70 (must be cautious not to extrapolate out of this range where this equation may not be valid anymore). Thus for a 45 year old person, the on-the-average SBP is 115.706 + 1.105 * 45 = 165.431 mmHg.

ASSUMPTIONS FOR THE LINEAR REGRESSION MODEL - RESIDUAL ANALYSIS

The **residue** of each observation is given by the difference between the observed value and the fitted value of the regression line. For example, from the dataset, we have a 50 year-old person with systolic BP of 164 but the fitted-value from the regression line is 168.3 (see Fig. 2). Thus the residue for this person is -4.3 (164 - 168.4). For this dataset, we will have 55 residual points.

For the linear regression model to be valid, there are three assumptions to be checked on the residues:

a. No outliers.
b. The data points must be independent.
c. The distribution of these residuals should be normal with mean = 0 and a constant variance.

a. Checking outliers

In template II, tick on the **Casewise Diagnostic** box (default value of three standard deviations should be fine) and table IIIa is obtained.

Table IIIa

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	158.8004	189.2821	171.5091	8.74338	55
Residual	-15.1799	18.2799	.0000	9.01606	55
Std. Predicted Value	-1.454	2.033	.000	1.000	55
Std. Residual	-1.668	2.009	.000	.991	55

[a] Dependent variable: Systolic blood pressure (mmHg).

Our interest is in the Std (Standardised) Residual; making sure that the minimum and maximum values do not exceed ±3. Here, we do not have any outliers.

b. Checking independence

In template II, tick the Durbin-Watson box to have this estimate included in the model summary (see Table IIIb)

Table IIIb. Durbin-Watson Estimate

Model summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.696[a]	.485	.475	9.10072	2.530

[a] Predictors: (Constant), Age (years).

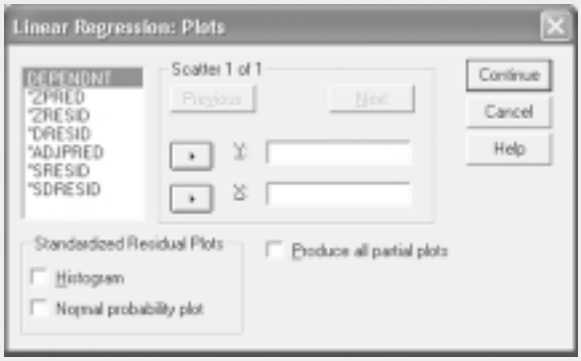
[b] Dependent variable: Systolic blood pressure (mmHg).

The Durbin-Watson estimate ranges from zero to four. Values hovering around two showed that the data points were independent. Values near zero means strong positive correlations and four indicates strong negative. Here, the independence assumption is satisfied.

c. checking the normality assumptions of the residuals

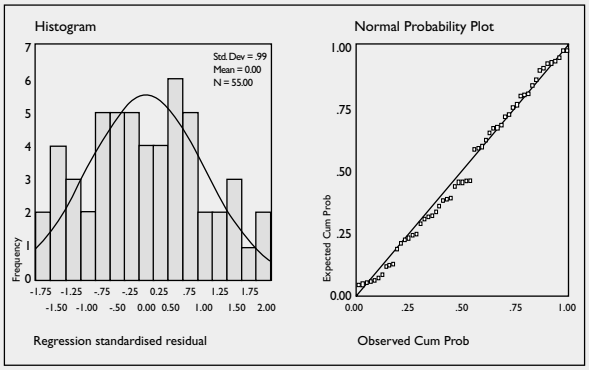
In template I, click on the **Plots folder** to get Template IV.

Template IV



Tick on the Histogram and Normal probability plot to get Fig. III to check on the normality assumptions of the residues.

Fig. III Histogram and Normal Probability Plot.

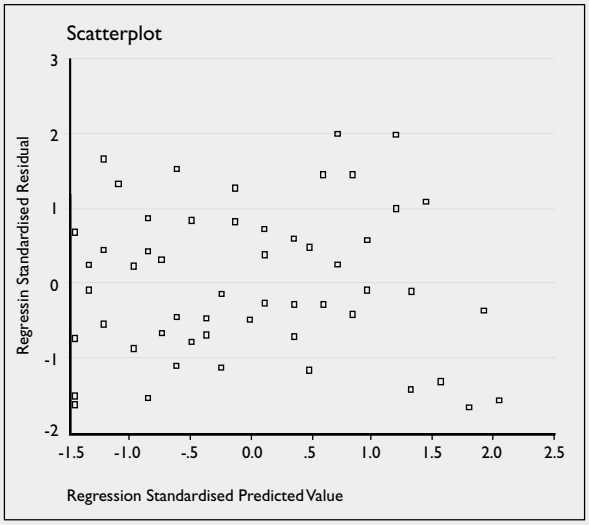


The distribution of the residual satisfies the normality assumptions[2].

d. Checking for constant variance

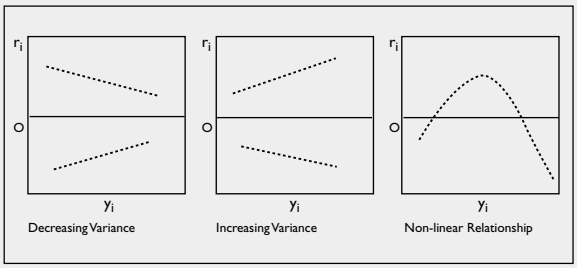
In template IV, select ***ZRESID** (Regression Standardized Residual) into the Y box and ***ZPRED** (Regression Standardized Predicted Value) into the X box to get Fig. IV :

Fig. IV Scatter plot of standardized residual vs standardised predicted value.



What do we want to see? As long as the scatter of the points shows no clear pattern, then we can conclude that the variance is constant. See Fig. V for problematic scatter plots.

Fig.V Problematic scatter plots.



MULTIPLE LINEAR REGRESSION

Given that we have also collected the smoking status of each subject, a multiple regression model with both age and smoking status correlating with systolic BP could be performed.

Since smoking status is a categorical variable, we need to understand the numerical coding, say, smoker = 1 & non-smoker = 0. In this case, when the multiple regression is performed, the regression estimate in the model for smoking status will be for the smoker comparing with the non-smoker, see Table IV.

Table IV. Multiple Regression model for Systolic BP with age and smoking status.

Coefficients[a]

Model	Unstandardised Coefficients		Standardised Coefficients			95% Confidence Interval for B	
	B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1 (Constant)	110.667	7.311		15.136	.000	95.996	125.338
Age (years)	1.055	.134	.699	7.893	.000	.787	1.324
Smoker	8.274	2.234	.328	3.703	.001	3.791	12.758

[a] Dependent variable: Systolic blood pressure (mmHg).

How can we interpret the result?

1. An Adjusting for covariate/confounder model

If our interest is only to determine whether age affects systolic BP after taking into account the smoking status, from table IV, we say that age is still statistically significantly affecting systolic BP (and the p-value of the smoking status is of no interest).

2. A Predictor model

In this case, the p-values of all variables would be of interest. From table IV, we conclude that both

age and smoking status are significant risk factors affecting the systolic BP. A smoker has on the average 8.3 (95% CI 3.8 to 12.8) higher BP compared to a non-smoker (given the same age).

Which independent variable has more influence on SBP? This will be given by the (absolute) value of the Standardized Coefficients Beta, the bigger the more influence. In this example, Age (Beta = 0.699) has a heavier influence on Systolic BP than the Smoking status (Beta = 0.328). If we have collected the information of whether a subject exercised or not, then Beta for Exercise will be negative (since exercise have a negative effect on increase of SBP).

ADJUSTED R SQUARE

In multiple regression, the R measures the correlation between the observed value of the dependent variable and the predicted value based on the regression model. The sample estimate of R Square tends to be an overestimate of the population parameter; the Adjusted R Square is designed to compensate for the optimistic bias of R Square, see Table V.

Table V.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.770[a]	.592	.577	8.17306

[a] Predictors: (Constant), Smoker, Age (years).

Age alone explains only 48.5% of the variance on SBP and when including the Smoking status, this increases to 57.7%. As we include more independent variables in the model, the Adjusted R Square will 'improve'.

CATEGORICAL VARIABLES WITH MORE THAN TWO LEVELS

Usually Race has 4 levels (with coding 1 = Chinese, 2 = Indian, 3 = Malay & 4 = Others). We cannot simply put Race as one of the variables in the model for the coding is arbitrary and the regression estimate obtained for Race will not make sense. A reference category has to be chosen, lets' say Chinese, and we have to create **Dummy** variables for the rest of the races. Table VI shows the three new dummy variables for Indian, Malay and Others by using the **Recode** option.

Table VI. Dummy variables for Race.

Subject	Race	Indian	Malay	Others
1	1 (Chinese)	0	0	0
2	1 (Chinese)	0	0	0
3	3 (Malay)	0	1	0
4	2 (Indian)	1	0	0
5	4 (Others)	0	0	1
6	2 (Indian)	1	0	0

Table VII shows the regression estimates for the model with age, smoking and race. The way to interpret the 'Race' regression estimates will be 'the Indians on the average have 1.98 mmHg higher in systolic BP with the Malays and Others having lower systolic BP compared to the Chinese' but this is not statistically significant.

MULTI-COLLINEARITY

When multiple regression is applied in a situation where there are moderate to high intercorrelations among the independent variables, two situations may happen. Firstly, the importance of a given explanatory variable is difficult to be determined because the effects are confounded **(distorted p-values)** and the other is that **dubious relationships may be obtained**.

Table VII. Regression model with Age, Smoking status and Race.

Coefficients[a]

Model	Unstandardised Coefficients		Standardised Coefficients	t	Lower Sig.	95% Confidence Interval for B	
	Std. B	Error	Beta			Upper Bound	Bound
1 (Constant)	112.449	7.648		14.704	.000	97.080	127.818
Age (years)	1.033	.136	.684	7.590	.000	.760	1.307
Smoker	7.758	2.300	.307	3.373	.001	3.136	12.379
INDIAN	1.977	2.871	.067	.689	.494	-3.793	7.747
MALAY	-1.861	3.177	-.058	-.586	.561	-8.245	4.523
OTHERS	-2.742	3.246	-.082	-.845	.402	-9.265	3.781

[a] Dependent variable: Systolic blood pressure (mmHg).

Table VIII shows the correlation between age, weight and height of the 55 subjects.

Table VIII

Correlations

		Age (years)	weight (kg)	height (m)
Age (years)	Pearson Correlation	1	.005	.840**
	Sig. (2-tailed)	.	.968	.000
	N	55	55	55
Weight (kg)	Pearson Correlation	.005	1	.547**
	Sig. (2-tailed)	.968	.	.000
	N	55	55	55
Height (m)	Pearson Correlation	.840**	.547**	1
	Sig. (2-tailed)	.000	.000	.
	N	55	55	55

** Correlation is significant at the 0.01 level (2-tailed).

There are significant moderate to high correlations between Height with Age and Weight. What happens when we perform a multiple regression model?

Table IX. Multiple Regression Model with Multicolinearity.

Coefficients[a]

	Unstandardised Coefficients		Standardised Coefficients		
Model	B	Std. Error	Beta	t	Sig.
1 (Constant)	-87.123	2987.566		-.029	.977
Age (years)	.829	3.098	.549	.268	.790
Smoker	7.517	2.542	.298	2.957	.005
INDIAN	1.747	3.071	.060	.569	.572
MALAY	-2.525	4.271	-.078	-.591	.557
OTHERS	-3.362	4.265	-.100	-.788	.434
weight (kg)	-3.126	3.107	-.054	-.040	.968
height (m)	6.661	2.577	.159	.065	.948

[a] Dependent variable: Systolic blood pressure (mmHg).

We can observe that the p-value of Age has become not significant and a dubious-negative relationship between weight and SBP is obtained, see Table IX. Another tell-tale sign of multicolinearity is that the Adjusted R Square is severely reduced as the explanatory variables are largely attempting to explain much of the same variance in the response variable.

Pearson's correlation only enable us to check multicolinearity between any two variables; but sometimes a variable could be co-linear with a combination of other variables. In this case, we can use the **tolerance measure** which gives the strength of the linear relationships among the independent variables.

To get this measure, in Template II, tick on the **Collinearity diagnostic** box to get Table X.

Tolerance lies between zero to one (the VIF is just the reciprocal of tolerance). A value close to zero indicates that a variable is almost a linear combination of the other independent variables. From Table X, Age, Weight & Height were **multicollinear**.

What's an acceptable tolerance range? Values above 0.6 would be recommended but since most likely there will be some correlation between variables (especially with dummy variables), 0.4 and above would be acceptable.

One way to combat the above issue is to combine explanatory variables that are highly correlated (e.g. taking their sum). An alternative is simply to select one of the set of correlated variables for use in the regression analysis.

Let's say we remove Height (since lowest tolerance) from the model.

Table X. Multiple Regression Model with Tolerance Measures.

Coefficients[a]

	Unstandardised Coefficients		Standardised Coefficients			Collinearity Statistics	
Model	B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1 (Constant)	-87.123	2987.566		-.029	.977		
Age (years)	.829	3.098	.549	.268	.790	.002	506.538
Smoker	7.517	2.542	.298	2.957	.005	.819	1.221
INDIAN	1.747	3.071	.060	.569	.572	.756	1.323
MALAY	-2.525	4.271	-.078	-.591	.557	.474	2.108
OTHERS	-3.362	4.265	-.100	-.788	.434	.517	1.934
Weight (kg)	-3.126	3.107	-.054	-.040	.968	.005	216.734
Height (m)	6.661	2.577	.159	.065	.948	.001	723.602

[a] Dependent variable: Systolic blood pressure (mmHg).

Table XI

Coefficients[a]

Model	Unstandardised Coefficients		Standardised Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1 (Constant)	106.890	23.992		4.455	.000		
Age (years)	1.030	.138	.682	7.445	.000	.970	1.031
Smoker	7.522	2.514	.298	2.992	.004	.819	1.220
INDIAN	1.769	3.021	.060	.586	.561	.765	1.307
MALAY	-2.533	4.224	-.079	-.600	.551	.475	2.106
OTHERS	-3.387	4.204	-.101	-.806	.424	.521	1.919
Weight (kg)	.075	.307	.032	.245	.808	.464	2.156

[a] Dependent variable: Systolic blood pressure (mmHg).

This model is now statistically 'stable'.

MODEL SELECTIONS

The above models have been based on the Enter option which included all the independent variables into the model regardless of their significance.

Template V. Model Selection Options.

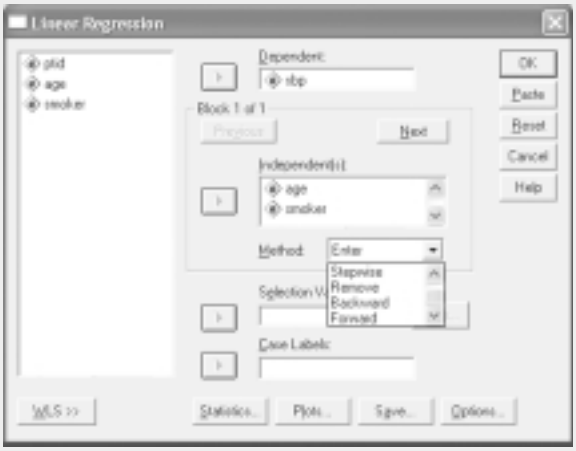


Table V shows the various model selection options available.

a. Forward
This model selection starts to include variables by their order of significance. Only variables that have $p < 0.05$ are in the model. This method is usually used in an exploratory study where one is not so sure what are the important variables influencing the outcome.

b. Backward
This method starts with all the variables in the model and variables are excluded on the basis of their non-significance. Usually used for a confirmatory study on the important variables influencing the outcome.

c. Stepwise/Remove
This is the combination of the forward and backward methods. In the stepwise method, variables that are entered will be checked at each step for removal. Likewise, in the removal method, variables that are excluded will be checked for re-entry.

How should we then derive our models? Multicollinearity should be carried out first before we perform the above model selections and then the checking of the residual-assumptions for the derived model to be done before we can 'accept' it as the final model.

To conclude, the material covered here only highlighted the basic and essential understanding of Linear Regression Analysis; you are encouraged to do further reading[5-9]. Our next article, Biostatistics 202 : Logistic Regression Analysis, will discuss on how to analyse the situation when the outcome variable is categorical.

REFERENCES

1. Chan YH. Biostatistics 101: Data Presentation. Singapore Med J 2003; 44:280-5.
2. Chan YH. Biostatistics 102: Quantitative Data – Parametric and Non-parametric tests. Singapore Med J 2003; 44:391-6.
3. Chan YH. Biostatistics 103: Qualitative Data – Tests of Independence. Singapore Med J 2003; 44:498-503.
4. Chan YH. Biostatistics 104: Correlational Analysis. Singapore Med J 2003; 44:614-9.
5. Kleinbaum DG, Kupper LL, Muller KE, Nizam A. Applied Regression Analysis and Other Multivariable Methods. Pacific Grove, CA:Brooks/ Cole, 1998.
6. Lewis-Beck MS. Regression Analysis, Beverley Hills, CA:Sage, 1993.
7. Wayne DW. Biostatistics. 6th ed. New York: John Wiley & Sons, 1995.
8. Draper NR, Smith H. Applied Regression Analysis. 2nd ed. New York: John Wiley & Sons, 1981.
9. Neter J, Kutner MH, Nachtsheim CJ, Wasserman W. Applied Linear Regression Models. 3rd ed. Chicago: Irwin, 1996.