chapter 2-6

Simple Linear Regression & Multiple Linear Regression

Introduction

In the previous chapter the correlation coefficient was discussed as a measure of association between two variables. The next step is to determine the equation of the best fitting straight line through the data, a process called Linear Regression analysis. Linear regression analysis allows you to find out how well you can predict one variable (dependent) from another (independent) variable. With multiple regression there is more than one independent variable used in the equation (note that in this case, the variables may not be completely independent from each other). As well as serving a predictive function, multiple regression allows for adjustment for the effects of other independent variables (also called confounders). The correlation coefficient is generated in the analysis, as discussed earlier is the measure of the association between variables, but it does not tell how well the equation can predict the dependent variable. The ability to predict is determined by the size of the **standard error of estimate** (S.E.E.). The calculation and interpretation of the S.E.E. will be discussed later.

Linear Regression

Linear regression analysis provides us with the best fitting straight line ($Y = b_0 + b_1X$, where b_1 = slope and b_o = intercept) through our data points. The *Y* variable is the one that is being predicted and is referred to as the dependent variable. The *X* variable is the one being used to make the prediction and is referred to as the independent variable, (or explanatory or predictor variable). The analysis provides the best estimates for b_0 *and* b_1.

Figure 2-6.1 is a diagrammatic representation of how the best fitting regression line is calculated. The best fitting line is determined by consideration of the **deviations** between the **observed** data point and the **predicted** value of *Y* for all given values of *X*. In the diagram, the vertical distance labeled d is the error or difference between predicted and observed data. Differences (*d*) are calculated for all data points. These values of *d* are then squared and summed ($\sum d^2$ - Sum of squared deviations). The best fitting line through the data points is defined as the line that has the smallest or least sum of these squared deviations. Hence this method is called “least sum of squares curve fitting.” These *d*’s are called residuals (observed - predicted) and their importance will be discussed later, particularly in chapter 2-12, where residual analysis will be used as part of the modeling process.

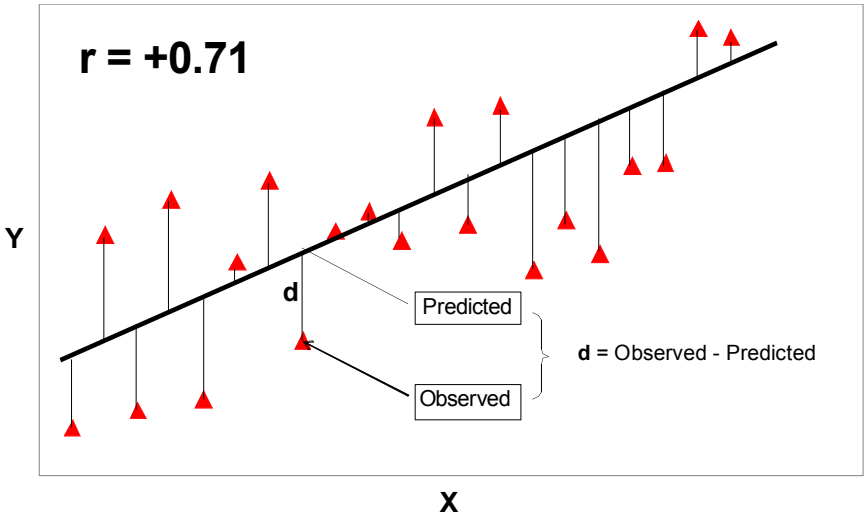


Figure 2-6.1: Linear Regression using least sum of squares line fitting.

The basic equation for the prediction of the dependent variable (*Y'*) from the independent variable (*X*), requires the calculation of a slope (b_1) and intercept (b_0).

$$Y' = b_1 X + b_0$$

Figure 2-6.2 shows the data on right and left hand lengths from university men used to demonstrate the calculation of the correlation coefficient, r, in the previous chapter in Table 2-5.1. It is now used to illustrate the calculation of the regression coefficients for the equation of left hand length predicting right hand length.

Therefore, in this example shown in Figure 2-6.2, the dependent variable (*Y*) is the right side hand length, which is predicted by the independent variable (*X*), left side hand length. The equations for the calculation of the regression coefficients more commonly termed the slope (*m or* b_1) and the intercept (*c or* b_0) are:

$$b_1 = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n(X^2) - (\sum X)^2}$$

$$b_0 = \overline{Y} - b_1 \overline{X}$$

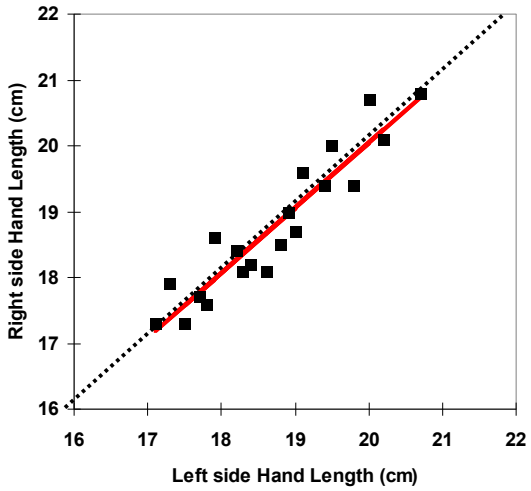


Figure 2-6.2: Regression line of Left side Hand Length predicting Right side Hand Length in university men (n=20). Dotted line is the line of identity.

Hand Length Right Side	**Hand Length Left Side**		
Y	X	XY	X^2
18.6	17.9	332.94	320.41
17.9	17.3	309.67	299.29
19.4	19.4	376.36	376.36
18.1	18.6	336.66	345.96
17.3	17.1	295.83	292.41
17.6	17.8	313.28	316.84
18.4	18.2	334.88	331.24
17.7	17.7	313.29	313.29
20.1	20.2	406.02	408.04
20	19.5	390	380.25
18.1	18.3	331.23	334.89
20.7	20	414	400
19.4	19.8	384.12	392.04
19.6	19.1	374.36	364.81
20.8	20.7	430.56	428.49
18.5	18.8	347.8	353.44
17.3	17.5	302.75	306.25
18.2	18.4	334.88	338.56
18.7	19	355.3	361
19	18.9	359.1	357.21
$\sum Y$	$\sum X$	$\sum XY$	$\sum X^2$
375.4	374.2	7043.03	7020.78
$\overline{Y}$	$\overline{X}$		
18.77	18.71		
***b*$_1$** (slope)	$b_1 = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n(X^2) - (\sum X)^2}$ $b_1 = \frac{20(7043.03) - (374.2)(375.4)}{20(7020.78)(374.2^2)} = 0.99$		
b$_0$ (intercept)	$c = \overline{Y} - m\overline{X}$ $b_0 = 18.77 - 0.99(18.71) = 0.254$		
Regression Equation	**Right Hand L. = 0.99 Left Hand L. + 0.254** **r = 0.94 S.E.E. = 0.38cm**		

Table 2-6.1: Calculation of Regression Coefficients for the prediction of Right Hand Length from Left Hand Length in university men (n=20).

Using these equations, it was found that the equation Right Hand Length = 0.99 Left Hand Length + 0.254 was the best fitting straight line to predict Right Hand Length from Left Hand Length (Table 2-6.1). Figure 2-6.2 shows this line plotted through the data points. The points seem quite uniformly scattered around the line, showing that a straight line is a good description of the relationship. The dotted line is the line of identity, where $b_1 = 1$ and $b_0 = 0$. Since in the analysis *m* was found to be 0.99, the two lines are virtually coincident, although the regression line is shifted slightly upwards by the intercept of 0.254cm, inferring that right hands are slightly longer than left hands. It should be noted that despite this shift, paired t-test analysis showed that there was in fact no significant difference in right and left hand lengths in this data.

The results of the regression analysis give us two important statistics in addition to the slope and intercept of the best fitting straight line through the data. The first one is the correlation coefficient (r) along with its associated probability level. With respect to the linear regression analysis for the slope and intercept to have any meaning, there must first be a significant ($p<0.05$) correlation coefficient indicating that relationship exists. The correlation coefficient quantifies the degree of association between the two variables and the interpretation of r was discussed in chapter 2-5. Unfortunately, many people will use the correlation coefficient as their indicator of how well the equation can predict. This is wrong! The r tells you the degree of association, not how well the equation can predict.

Standard Error of Estimate

The statistic that does tell you how well the equation predicts is the Standard Error of Estimate (S.E.E.). The S.E.E. describes the variability about the line with respect to the dependent variable Y.

In linear regression there are three main assumptions made about the relationship between Y and X with respect to the variability of Y about the line, illustrated in Figure 2-6.3:

1. For any value of X, there is a normal distribution of Y values from which the sample value of Y is drawn.
2. For any given value of X, the corresponding population of Y values has a mean of μ that lies on the straight line $\mu = \alpha + \beta(X - X) = \alpha + \beta x$, where α and β are parameters.
3. In each population, the standard deviation of Y about its mean has the same value, often denoted by $\sigma_{y'x'}$ This is referred to as homoscedasticity. If the variability of Y about the line varied with different values of X then this would be termed heteroscedasticity.

The S.E.E. is the standard deviation of this normal distribution of Y about the regression line and therefore has the same units as Y. If you remember back to the properties of the normal distribution, 68.26% of scores lie between -1 and +1 standard deviations from the mean. This

can be applied to interpretation of the S.E.E. in that when predicting with a regression equation we will be within plus or minus one S.E.E. of the true score 68.26% of the time. It is the S.E.E. therefore, and not the correlation coefficient, that tells you how well an equation can predict. There are no tables that tell you how "good" a S.E.E. should be. Each calculated S.E.E. must be evaluated in relation to the application required for the equation.

Table 2-6.2 shows the results of regression analysis on our samples of university men and women in order to produce regression equations to predict Standing Height from Tibial Height (Knee Height). The rationale for this type of equation is that it could be used to estimate height in individuals who are confined to a wheelchair. Are the prediction equations good enough? Both equations are significant (p<0.05), with correlation coefficients of 0.79 and 0.84 for men and women, respectively.

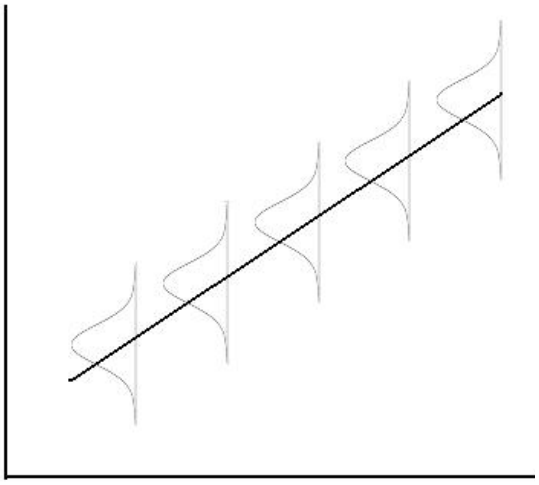


Figure 2-6.3: Distribution of error about a regression line

The S.E.E.s however, are 3.89 cm for men and 4.27 cm for women. Thus, height can be predicted with an error of approximately plus or minus 4cm, 2 out of 3 times (68.26%). The decision as to whether this is good enough lies solely with the user of the equation. Is this margin of error small enough for the purpose required of the equation? Possibly, it depends upon the situation. Interestingly, the equation for women has the poorer S.E.E., yet the higher correlation coefficient, which further illustrates how r can be misleading.

Another example is the prediction of %body fat from skinfold measures. The S.E.E. of these equations is in the order of 3.7% of body fat (Jackson & Pollack 1985), indicating that when the equation is used to predict the body fat of an individual, the prediction in approximately 2 out of 3 (actually 68.3%) times will be within plus or minus 3.7% body fat of the correct value. Thus, if a prediction of 15% body fat is made then the confidence in that prediction would be that on 2 out of 3 occasions, the body fat actually lies between 11.3 - 18.7% body fat of the correct value. Obviously the usefulness of a methodology carrying this degree of error is limited in individual assessments. So be wary of regression equations being reported only with r. Find out the

Sex	b_1	b_0	r	S.E.E.	*p*
Male	1.78	10.74	0.79	3.89cm	0.00
Female	1.74	12.55	0.84	4.27cm	0.00

Table 2-6.2: Linear regression of Tibial height predicting Height in University Males (N=49) and Females (N=67)

S.E.E. for any equation you use to predict with, and satisfy yourself that it is good enough for the use you have for it.

Multiple Regression

Multiple regression is an extension of linear regression where more than one independent variable is used. Figure 2-6.4 is a conceptualization of how multiple regression works. In this venn diagram, the circles represent the variance of the four variables. As discussed previously (Chapter 2-5), the degree of overlap of the circles represents the percentage of variance explained as quantified by R^2 (the coefficient of determination). In this example, shown in figure 2-6.4, X_1 has the highest correlation with Y, therefore it would be the first variable included in the regression equation based upon least sum of squares fitting. The next question is which is the next best independent variable to add into a multiple regression equation? X_3 has a higher correlation with Y than X_2,; however, X_2 would be a better choice than X_3 to include in an equation with X_1 to predict Y. Although, X_2 has a lower correlation with Y than X_3, in combination with X_1 it explains more of the variance in Y than the X_1 X_3 combination.

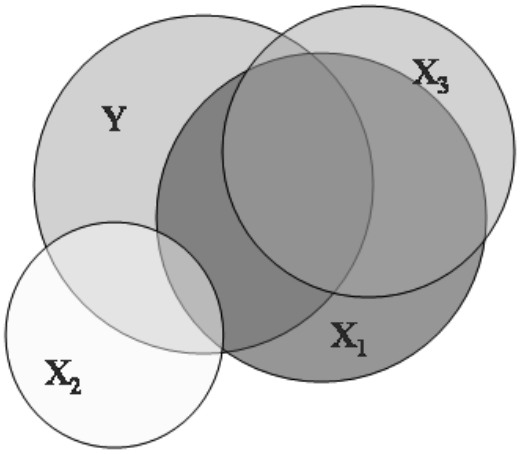


Figure 2-6.4: Venn diagram illustrating explanation of variance in dependent variable (Y) by 3 independent variables (X_1, X_2, X_3).

The model of the multiple regression looks like:

$$\hat{Y} = b_o + b_1X_1 + b_2X_2 + b_3X_3 \ldots\ldots b_kX_k$$

The coefficient b_1 represents the unit change in Y per unit change in X_1 taking into account the association between X_2 and Y etc. and is referred to as a partial regression coefficient. There are as many regression coefficients (b_k) as there are independent variables. The regression coefficients are estimated using the criterion of least sum of squares. These coefficients are called unstandardized regression coefficients. The magnitude of these coefficients does not tell us directly how predictive the variable is because the units of the independent variables might be very different. If however, you convert all the variables into standard scores (mean = 0, s.d. = 1) and then run the regression you produce standardized regression coefficients or beta weights. The resulting coefficients can be then compared directly, and give relative importance of the variable.

Multiple Regression Predictor Variables

The following examples show regression of continuous and binary predictor variables on a continuous dependent variable as illustrated in a study of the forced expiratory volume (amount of air breathed out in one second, FEV_1) of children aged 7 to 11 years. As shown in the output in Table 2-6.3, the regression equation is:

FEV_1= -2.2075 + 0.0853 Age + 0.0246 Height.

Note that because of the additive nature of the equation, the regression coefficients are smaller than they would be if age alone, or height alone, were modelled in a simple linear regression equation. The *t* statistics and corresponding *P*-values for age and height test the null hypotheses that there is no association of FEV_1 with age (or height) after controlling for its association with height (or age). In this case the null hypothesis was rejected at the level of <0.001.

FEV1	Coeff. b	Std Error b	Beta	*t*	*P>t*
Age	0.0853	0.0154	0.8801	5.607	0.000
Height	0.0246	0.0016	0.7945	13.77	0.000
(Constant)	-2.2075	0.1811		-12.632	0.000

Table 2-6.3: Predicting FEV_1(litres/sec) in children aged 7 to 11 from age (yrs) and height (cm)

Table 2-6.4 shows the analysis of variance table which shows how the joint effects of age and height explain the variation in FEV_1. Sum of squares are divided into two components:

- *Sum of squares due to the regression of FEV_1 on both age and height*
- *Residuals Sum of Squares*

Source of Var.	SS	d.f	MS	F	p
Regression	25.6383	2	12.8192	244.3	0.0000
Residual	33.2201	633	0.05248		
Total	58.8584	635	0.09269		

Table 2-6.4: Analysis of variance table for the predicting of FEV_1 (litres/sec) in children aged 7 to 11 from age (yrs) and height (cm)

The 2 degrees of freedom are due to the two independent variables. The mean square is calculated as the corresponding sum of squares divided by the degrees of freedom (25.6383 / 2 = 12.8192). The F statistic is calculated by dividing MS regression by the MS residuals (12.8192 / 0.05248 = 244.3). The square root of the residual mean square (MS residual) is the S.E.E. for the multiple regression. Therefore $S.E.E. = \sqrt{0.05248} = 0.229 \text{ litres/sec}$. The coefficient of

determination R^2 is the proportion of the total variability in Y attributable to the dependence of Y on all the X_i as defined by the regression model fit to the data, and in this example equals SS regression/SS total) = 25.6383/58.8584 = 0.4356. This indicates that the regression accounts for 43.56% of the total variance in FEV_1. The multiple correlation coefficient *R* therefore is equal to $\sqrt{.4356} = 0.66$.

Indicator Variables

An indicator variable is binary and coded as 0 or 1. Generally 0 indicates a lack of the characteristic or is the reference condition. The code of 1 is used to indicate that an individual has the specific characteristic. An example is sex, with 1 = female and 0 = male. In this case the regression coefficient for the indicator variable is the difference between the mean in girls to the mean of boys. (Note that if there was a regression only with this indicator variable the *t* statistic and corresponding *P*-value would be the same as derived from a t-test.). Using the previous example, if we added the indicator variable of sex to the previous equation, then the regression coefficient for the variable sex estimates the difference in mean FEV_1 in girls compared to boys, having allowed for the effects of age and height.

Figure 2-6.5 illustrates the interaction between a binary variable “Asthmatic Status”, and a continuous independent variable Height. The relationship of height of children to lung function (deadspace) measurements takes into account their asthmatic status. In this case, whether the person has asthma or not makes an important difference in the relationship of height to deadspace. Predictor variables (one binary and one continuous) can be tested for interaction by creating a multiple of the two variables (in this case asthmatic status x height).

Where an indicator variable has more than 2 categories (e.g. age groups of 1-4, 5-9, 10-14, 15-19; level of exposure as low, medium and high) then dummy variables must be used. A baseline group is chosen which is usually the lowest coded value) and dummy variables are created such that *k*-1 indicator variables are needed for *k* levels. One such example is the variable “Smoking Status” with categorical values of 0 = never; 1 = ex-smoker; 2 = current smoker. This can be recoded to new variables, labeled as “es” and “cs”. The dummy variable “es” has the value of 1 if smoking status = 1 (everything else = 0); cs = 1 if smoking status = 2 (everything else=0). The comparison group then is never smoker: (cs = 0 and es = 0), with comparisons made to es and cs, entered together in the regression analysis. Note that to interpret dummy variables so that they have the same comparison group, all dummy variables must be entered into the equation together. This is important when stepwise regression or other selection approaches are used, as will be described later.

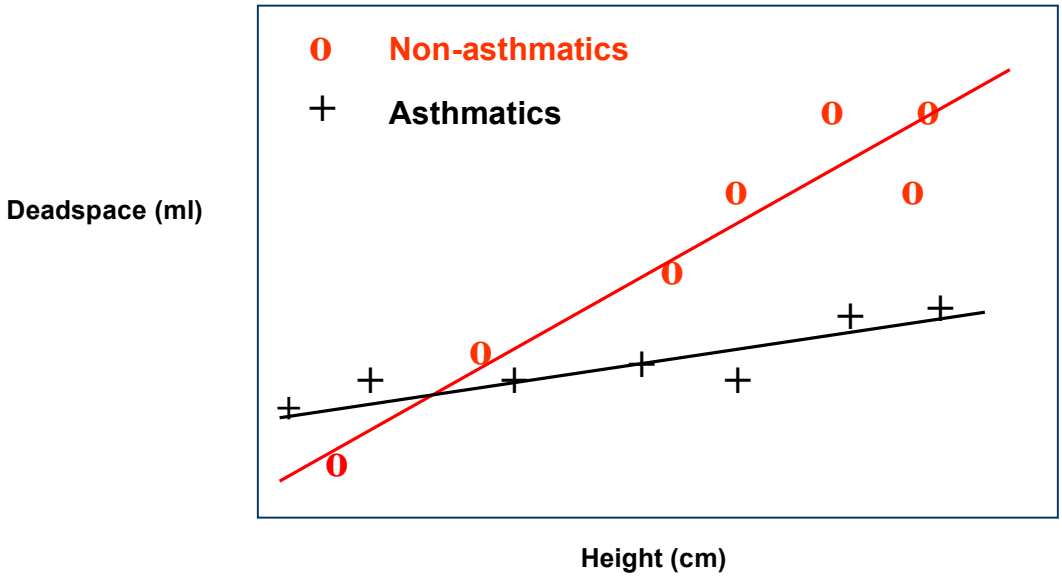


Figure 2-6.5: Regreesion analysis of Height predicting Deadspace for Non-asthmatic and Asmatic subjects

Figure 2-6.6 illustrates such an analysis. A 10% random sample was selected from the Canada Fitness Survey data, resulting in a sample of 1077 subjects. Waist to Hip Girth Ratio (Waist Girth / Hip Girth) is known to be related to age and sex and is also associated with smoking status of the individual. A regression analysis was carried out to predict Waist to Hip Girth Ratio from Age (years), Sex (Male = 1, Female = 2), an ExSmoker Dummy Variable (1 = ExSmoker, 0 = Others), a Current Smoker Dummy Variable (1 = Smoker, 0 = Others). An Age-Sex Interaction term calculated as Age multiplied by Sex (the ExSmoker and Smoker Dummy variables were calculated from a Smoking Status variable coded as 1 = Never Smoked, 2 = Current Smoker and 3 = Quit Smoking).

Using the $p<0.05$ criterion, only the ExSmoker Dummy variable was found to be not statistically significant. The conclusions therefore are (1) there is a relationship of Waist to Hip Girth Ratio with Age, and (2) that this relationship is different for the two sexes and between Smokers and NonSmokers. However, there is no difference in relationship for ExSmokers. This is not an unexpected finding in that an ExSmoker is defined in this study as someone who has quit smoking without information on the duration prior to the study. The interaction term for sex and age being significant indicates that the relationship of age to waist to hip girth ratio differs for men and women. If it were of interest, interaction terms for both of the smoking status dummy variables with age could have been created and evaluated in the same way.

Descriptive Statistics

	Mean	Std. Deviation	N
Wait to Hip Girth Ratio	.8307	.09005	1077
Age in years	38.364	13.4030	1077
Sex of Subject	1.546	.4981	1077
ExSmoker Dummy Variable	.2145	.41066	1077
Current Smoker Dummy Variable	.4011	.49035	1077
Age Sex Interaction	59.6760	29.81264	1077

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.749(a)	.561	.559	.05979

a Predictors: (Constant), Age Sex Interaction, ExSmoker Dummy Variable, Current Smoker Dummy Variable, Sex of Subject, Age in years

ANOVA (b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4.897	5	.979	273.962	.000(a)
	Residual	3.828	1071	.004		
	Total	8.725	1076			

(a) Predictors: (Constant), Age Sex Interaction, ExSmoker Dummy Variable, Current Smoker Dummy Variable, Sex of Subject, Age in years

(b) Dependent Variable: Wait to Hip Girth Ratio

Coefficients (a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta	B	Std. Error
1	(Constant)	.880	.018		48.273	.000
	Age in years	.004	.000	.540	7.971	.000
	Sex of Subject	-.085	.011	-.468	-7.596	.000
	ExSmoker Dummy Variable	.006	.005	.026	1.115	.265
	Current Smoker Dummy Variable	.011	.004	.060	2.627	.009
	Age-Sex Interaction	-.001	.000	-.350	-3.818	.000

(a) Dependent Variable: Waist to Hip Girth Ratio

Figure 2-6.6: Regression analysis of Waist to Hip Girth Ratio predicted from Age (yrs), Sex (Male = 1, Female = 2), ExSmoker Dummy Variable (1 = ExSmoker, 0 = Others), Current Smoker Dummy Variable (1 = Smoker, 0 = Others), Age-Sex Interaction (Age x Sex).

Procedures for Selecting Variables

Typically there are many variables to choose from when attempting to create the best prediction equation. There are three common approaches for selecting the best predictors. These include forward selection, backward selection, and stepwise selection. All three methods are generally considered as forms of stepwise regression.

Forward selection – This approach begins with a simple regression model. The first variable entered has the largest positive or negative correlation with the dependent variable, followed by one with the next largest partial correlation. The default criterion for selection is the probability of F-to-enter is 0.05. This method is not a typical choice due to theoretical difficulties.

Backwards elimination - This is the opposite approach in that all variables are entered first and then sequentially removed (probability of F-to-remove = 0.10).

Stepwise selection – This is essentially a combination of forward and backwards selection. In this process the best linear regression equation using one independent variable is determined. Then each other possible independent variable is tested to find out which, in combination with the first included independent variable, contributes most to the explanation in the variance of the dependent variable. The equation with these two independent variables is now produced with its associated r and S.E.E. The process is then repeated with a third independent variable, and on to the fourth etc. Each time a regression equation is produced with one more independent variable. Critical to this decision is the S.E.E. The S.E.E. may drop dramatically over the inclusion of the first 3 variables but then differ little with further inclusions. The choice here would be to use the equation with three independents, since the inclusion of more variables adds little to the predictive power.

The argument against using stepwise regression is the choices are made by a computer. In each successive step the next independent variable is chosen based upon F tests of all available variables. The argument is that when so many F tests are carried out, some of the variables included pass their F test purely by chance and that this would then be a result that could never be repeated in another sample. To guard against this, one should be careful about which variables are put into the inclusion list to be selected from by stepwise regression. Do not include variables that are not intuitively appropriate as predictors. In addition you might want to pare down variables to those that make practical sense. You might produce an equation that predicts a dependent score from anthropometric measures. It might seem inappropriate to have users of your equation, trained to use and provided with skinfold callipers. In this case it would make sense not to put skinfolds in the inclusion list, even though they might make good predictors for your equation. Be wary of using too many independent variables. Such equations are clumsy.

Pitfalls in the Use of Regression Equations

Large Sample Size: As with all statistics with very large sample sizes, statistical significance loses all relationship to clinical significance. Even very weak relationships (small r) and very poorly predicting equations (high SEE) can be statistically significant, such that statistical significance has little value in the interpretation of a regression equation.

Sample Size vs Number of Independent Variables: Be careful when carrying out multiple regression on small sample sizes. By including many independent variables, it is easy to obtain correlations that are very large but could not be replicated in another sample. As a general rule the sample size should be at least 5 times the number of independent variables you use in the equation.

Restricted by the Range of Data: Regression equations are fit to the data in the sample, therefore they are only justified within the range of that data. You have no knowledge of the relationship outside of that; therefore, if you use the equation on data outside of that range you have no measure of confidence about the predicted values. Regression equations should not be used outside the range of the originating data.

Sample Specificity: The regression equation will fit best on the data it was originated on. It can never perform as well on another data set. However, cross-validation studies entail predicting on a different sample in order to test how valid and sample specific the equation is.

Spurious Correlation: Using a predictor variable that is contained within the calculation of the dependent variable. E.g. predicting BMI from body weight when weight is in the calculation of BMI. You will get a spuriously high correlation between the two.

Multicollinearity: Another consideration is whether independent variables are highly correlated. An example is in time series or longitudinal data in which an exposure variable (e.g. cat ownership) is measured at age one and two years. As a result the parameter estimates will be correlated and the regression equation will have a significant R^2 and a low tolerance (defined as $1-R^2$) even though none of the coefficients are significantly different than zero.

Nonlinearity: Multiple regression generally is robust and slight deviations from the assumptions are not a problem. Non-linearity can occur in the relationship of two variables and there are a number of options.

1. The predictor variable can be categorized, typically in 2 to 5 groups instead, and analysis is done using dummy variables.
2. The predictor variable can be transformed, e.g. using logarithmic transformation. Other common transformations are X^2 and 1/X. If an X^2 term added to the X term is significant, this

would indicate nonlinearity.

3. Both the predictor and dependant variables can be transformed. A more common approach is quadratic transformation involving X^2 and Y^2. An example is the relationship of age^2 to height^2.

Model Sensitivity: This refers to how estimates are affected by a single data point or subgroups of the data. A large residual (difference between the observed and fitted data) is an outlier, and can be observed in a scatter plot. Plotting residuals against fitted values should show a similar spread with increasing fitted values (if it is increasing, one solution may be to log transform outcome Y). Leverage typically occurs when a data point is isolated far away from the cluster of points, but the regression line goes near or on it. An influential point has a large effect on the estimate (the Cook's D statistic can indicate this). With sensitivity analysis, the question posed is: do conclusions change if influential data are removed?

Linear Regression Analysis with MS EXCEL

In EXCEL, regression analysis can be carried out in two ways. Firstly, using paste functions, along with the CORREL function already mentioned, there are INTERCEPT, SLOPE and STEYX functions. This is useful if you wish to refer to these cells in equations entered in other cells.

The other option is to use the full regression analysis. Under the Tools menu you will find the Data Analysis option. Select this and you will be presented with a dialogue box where you can scroll down and select the regression package. If you select this option then you will see the dialogue screen shown in Figure 2-6.7. You must enter the location of the X and Y data in terms of the range of cell addresses in which they are located. This can be done by either pointing and dragging with the mouse or actually typing in the addresses. If you have labels at the top of the columns then there is a labels check box to select. You can ask for the line to be fit to have zero intercept. It sometimes makes biological sense to have a zero intercept, in which case select the "Constant is Zero" check box. Then select a cell for the top left hand corner of the summary report and enter it as

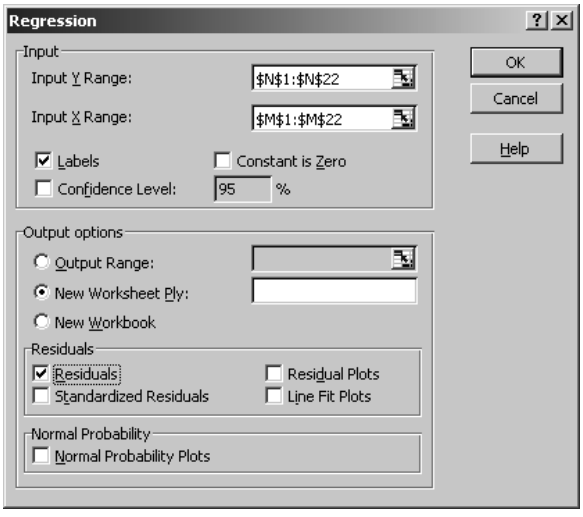


Figure 2-6.7: Regression Analysis dialog box

the “Output Range”. In addition several graphs can be selected. In this example the simple Line Fit Plot was selected. The resultant output is shown below:

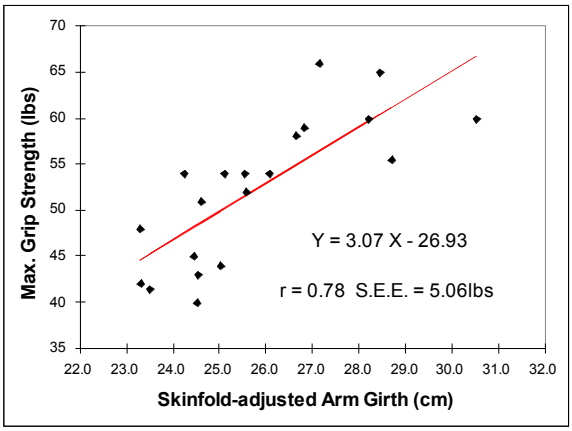


Figure 2-6.8: Plot of Regression Analysis of Skinfold-adjusted Arm Girth Predicting Maximum Grip Strengths in University Males

Figure 2-6.8 shows a plot of the Regression Analysis of Skinfold-adjusted Arm Girth Predicting Maximum Grip Strengths in a group of University men. Figure 2-6.9 shows the MS EXCEL output for this analysis. The red boxes highlight the most important parts of the results. The correlation coefficient = 0.78 (Multiple R – EXCEL labels it this way even though there might only be one independent variable). The SEE is 5.06lbs. The significance of F must be less than 0.05 for us to state that there is a significant relationship at the 95% level, which it is in this case.

SUMMARY OUTPUT	Males
Regression Statistics	
Multiple R	0.778533284
R Square	0.606114075
Adjusted R Square	0.584231523
Standard Error	5.057299904
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	708.4249182	708.4249182	27.69851026	5.27385E-05
Residual	18	460.3730818	25.57628232		
Total	19	1168.798			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-26.92995672	15.09863073	-1.78360258	0.091357898	-58.65102735	4.791113903	-58.65102735	4.791113903
SAFAGR	3.071244728	0.583560944	5.262937417	5.27385E-05	1.84522773	4.297261726	1.84522773	4.297261726

RESIDUAL OUTPUT

Observation	*Predicted gripR*	*Residuals*
1	49.85730396	-5.85730396
2	44.59933299	-2.499332987

Figure 2-6.9: Output for MS EXCEL Regression Analysis of Skinfold-adjusted Arm Girth Predicting Maximum Grip Strengths in Males (n = 20)

The intercept is -26.93 and slope for skinfold-adjusted forearm girth is 3.07. The equation is therefore:

Max. Grip Strength = 3.07 skinfold-adjusted forearm girth – 26.93

r = 0.78 SEE = 5.06lbs p<0.05

Regression Analysis with SPSS

Simple Linear Regression.

The linear regression analysis is found as the REGRESSION listing under the ANALYZE menu. Choose the LINEAR option and you will be presented with the dialog box shown in Figure 2-6.10. As with other SPSS dialog boxes the variables are sent over to selection boxes. In this case Grip strength was sent to the Dependent variable box and skinfold-adjusted forearm girth was sent to the independent(s) box. This will produce the simple linear regression. If a multiple regression equation is required, multiple independent variables are sent to the independent(s) box and the METHOD is changed to STEPWISE from ENTER. This will result in a series of equations being produced as discussed earlier. Figure 2-6.11 shows the SPSS output for the regression analysis called for in the dialog box in Figure 2-6.10. Since SPLIT FILE by SEX was on, there is a regression output for each of the sexes. The red boxes highlight the most important parts of the output: the r, SEE, significance, and the coefficients of the equation. Not surprisingly, the values for the men are exactly the same as the EXCEL output shown earlier. For the women the results are:

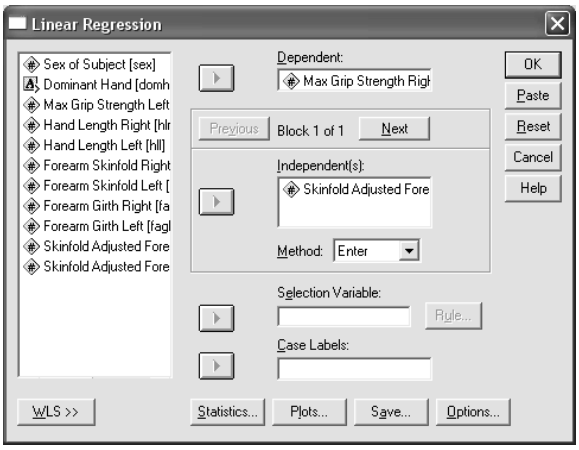


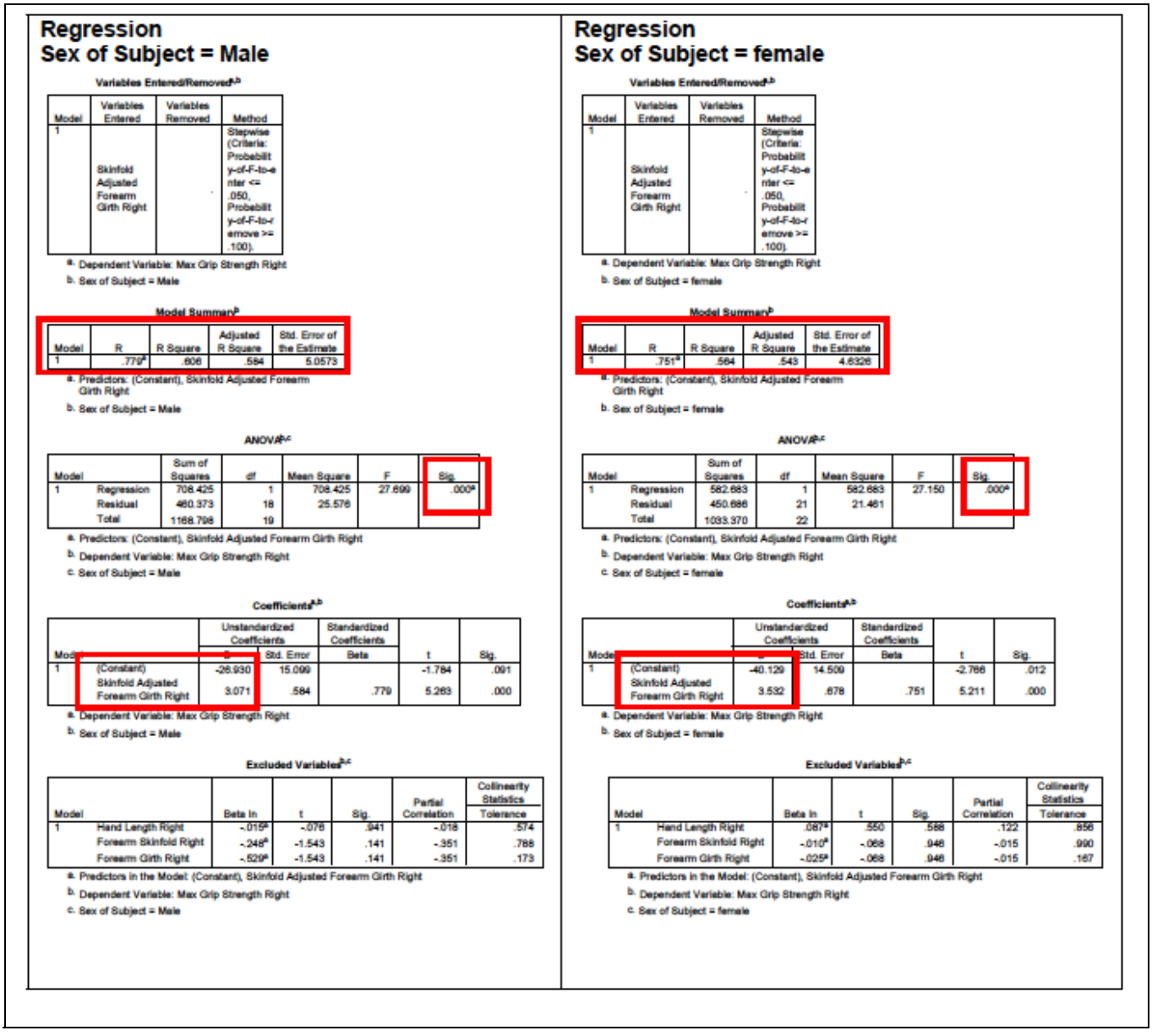
Figure 2-6.10: SPSS ANALYZE – REGRESSION – LINEAR – Dialog box for Regression Analysis of Skinfold-adjusted Arm Girth Predicting Maximum Grip Strengths in Males and Females Separately (Split File ON)

Max. Grip Strength = 3.53 skinfold-adjusted forearm girth – 40.13

r = 0.75 SEE = 4.63lbs p<0.05

Noticeably although the correlation coefficient is lower for the females the SEE is lower, which means the prediction equation for women is better than the prediction equation for men. An interesting question would be is it necessary to have separate equations for the two sexes. You could run a combined sex equation and see how the resultant equation would predict. However, if you wanted a test of the difference in the regression equations you would run an Analysis of

Covariance (ANCOVA). This will be dealt with in depth in the next chapter on tests of differences between means. The test would be an analysis of variance with Grip strength as the dependent, Sex as a grouping factor and skinfold-adjusted forearm girth as a covariate. If it was shown that Sex was not a significant factor then the two regression lines for the sexes are not significantly different from each other, and hence you do not need to have separate equations.



Regression
Sex of Subject = Male

Variables Entered/Removed[a,b]

Model	Variables Entered	Variables Removed	Method
1	Skinfold Adjusted Forearm Girth Right		Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: Max Grip Strength Right
b. Sex of Subject = Male

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.779[a]	.606	.584	5.0573

a. Predictors: (Constant), Skinfold Adjusted Forearm Girth Right
b. Sex of Subject = Male

ANOVA[a,c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	708.425	1	708.425	27.699	.000[a]
	Residual	460.373	18	25.576		
	Total	1168.798	19			

a. Predictors: (Constant), Skinfold Adjusted Forearm Girth Right
b. Dependent Variable: Max Grip Strength Right
c. Sex of Subject = Male

Coefficients[a,b]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-26.930	15.099		-1.784	.091
	Skinfold Adjusted Forearm Girth Right	3.071	.584	.779	5.263	.000

a. Dependent Variable: Max Grip Strength Right
b. Sex of Subject = Male

Excluded Variables[b,c]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance
1	Hand Length Right	-.015[a]	-.076	.941	-.018	.574
	Forearm Skinfold Right	-.248[a]	-1.543	.141	-.351	.788
	Forearm Girth Right	-.529[a]	-1.543	.141	-.351	.173

a. Predictors in the Model: (Constant), Skinfold Adjusted Forearm Girth Right
b. Dependent Variable: Max Grip Strength Right
c. Sex of Subject = Male

Regression
Sex of Subject = female

Variables Entered/Removed[a,b]

Model	Variables Entered	Variables Removed	Method
1	Skinfold Adjusted Forearm Girth Right		Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: Max Grip Strength Right
b. Sex of Subject = female

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.751[a]	.564	.543	4.6326

a. Predictors: (Constant), Skinfold Adjusted Forearm Girth Right
b. Sex of Subject = female

ANOVA[a,c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	582.683	1	582.683	27.150	.000[a]
	Residual	450.686	21	21.461		
	Total	1033.370	22			

a. Predictors: (Constant), Skinfold Adjusted Forearm Girth Right
b. Dependent Variable: Max Grip Strength Right
c. Sex of Subject = female

Coefficients[a,b]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-40.129	14.509		-2.766	.012
	Skinfold Adjusted Forearm Girth Right	3.532	.678	.751	5.211	.000

a. Dependent Variable: Max Grip Strength Right
b. Sex of Subject = female

Excluded Variables[b,c]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance
1	Hand Length Right	.087[a]	.550	.588	.122	.856
	Forearm Skinfold Right	-.010[a]	-.068	.946	-.015	.990
	Forearm Girth Right	-.025[a]	-.068	.946	-.015	.167

a. Predictors in the Model: (Constant), Skinfold Adjusted Forearm Girth Right
b. Dependent Variable: Max Grip Strength Right
c. Sex of Subject = female

Figure 2-6.11: SPSS Output for Regression Analysis of Skinfold-adjusted Arm Girth Predicting Maximum Grip Strength

Multiple Linear Regression.

Figure 2-6.6 showed the regression analysis of Waist to Hip Girth Ratio predicted from Age, Sex, ExSmoker Dummy Variable, Current Smoker Dummy Variable, and the Age-Sex Interaction (Age x Sex). This is an example of a multiple regression analysis and was achieved using the same dialog box as for the simple linear regression. The only difference being that a list of independent variables is provided for use in the analysis. In the case of the smoking analysis it was required that all the independent variables be entered simultaneously into one equation. Thus, the ENTER method was selected. However, the STEPWISE method would be selected if the researcher required a series of equations to be produced as described earlier and illustrated in Figure 2-6.12.

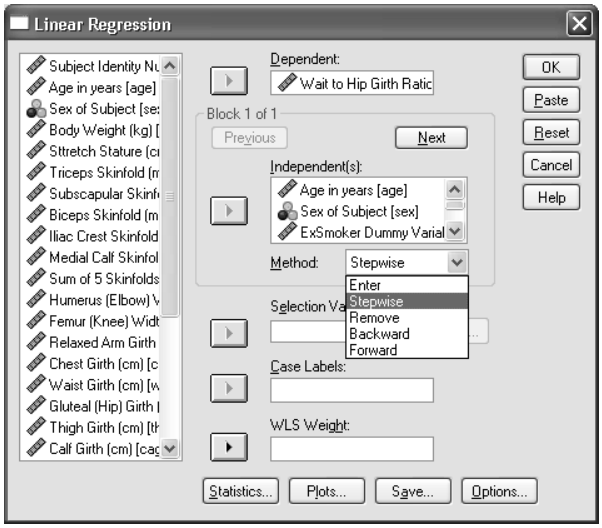


Figure 2-6.12: SPSS ANALYZE – REGRESSION – LINEAR – Dialog box for Multiple Regression Analysis of Waist to Hip Girth Ratio predicted from Age, Sex, ExSmoker Dummy Variable, Current Smoker Dummy Variable, and the Age-Sex Interaction (Age x Sex)

Reporting the Results of Regression Analysis

The relevant details to report following simple linear regression with one predictor variable are the regression equation, the correlation coefficient, the S.E.E., and the p-value (significance level). Following multiple linear regression (more than one predictor variable) report the regression equation, the coefficient of determination (R^2, total variance in the dependent variable that is explained by the independent variables), the S.E.E., and the p-values (significance levels) associated with predictor variables of interest.