How to Read the Output From Simple Linear Regression Analyses

This is the typical output produced from a simple linear regression of muscle strength (STRENGTH) on lean body mass (LBM). That is, lean body mass is being used to predict muscle strength.

Model Summary(b)

R	R Square	Adjusted R Square	Std. Error of the Estimate
.872(a)	.760	.756	19.0481
a Predictors: (Constant), LBM			
b Dependent Variable: STRENGTH			

ANOVA

Source	Sum of Squares	df	Mean Square	F	Sig.
Regression	68788.829	1	68788.829	189.590	.000
Residual	21769.768	60	362.829		
Total	90558.597	61			

Coefficients

Variable	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
(Constant)	-13.971	10.314		-1.355	.181	-34.602	6.660
LBM	3.016	.219	.872	13.769	.000	2.577	3.454

Table of Coefficients

The column labeled **Variable** should be self-explanatory. It contains the names of the items in the equation and labels each row of output.

The **Unstandardized coefficients (B)** are the regression coefficients. The regression equation is

STRENGTH = -13.971 + 3.016 LBM

The predicted muscle strength of someone with 40 kg of lean body mass is

-13.971 + 3.016 (40) = 106.669

For cross-sectional data like these, the regression coefficient for the predictor is the difference in response per unit difference in the predictor. For longitudinal data, the regression coefficient is the change in response per unit change in the predictor. Here, strength differs 3.016 units for every unit difference in lean body mass. The distinction between cross-sectional and longitudinal data is still important. These strength data are cross-sectional so differences in LBM and strength refer to differences between people. If we wanted to describe how an individual's muscle strength changes with lean body mass, we would have to measure strength and lean body mass as they change within people.

The **Standard Errors** are the standard errors of the regression coefficients. They can be used for hypothesis testing and constructing confidence intervals. For example, the standard error of the STRENGTH coefficient is 0.219. A 95% confidence interval for the regression coefficient for STRENGTH is constructed as ($3.016 \pm k\,0.219$), where k is the appropriate percentile of the t distribution with degrees of freedom equal to the Error DF from the ANOVA table. Here, the degrees of freedom is 60 and the multiplier is 2.00. Thus, the confidence interval is given by ($3.016 \pm 2.00\ (0.219)$). If the sample size were

huge, the error degress of freedom would be larger and the multiplier would become the familiar 1.96.

The **Standardized coefficients (Beta)** are what the regression coefficients would be if the model were fitted to standardized data, that is, if from each observation we subtracted the sample mean and then divided by the sample SD. People once thought this to be a good idea. It isn't, yet some packages continue to report them. Other packages like SAS do not. We will discuss them later when we discuss multiple regression.

The **t** statistic tests the hypothesis that a population regression coefficient β is 0, that is, H_0: $\beta = 0$. It is the ratio of the sample regression coefficient B to its standard error. The statistic has the form (estimate - hypothesized value) / SE. Since the hypothesized value is 0, the statistic reduces to Estimate/SE. If, for some reason, we wished to test the hypothesis that the coefficient for STRENGTH was 1.7, we could calculate the statistic (3.016-1.700)/0.219.

Sig. labels the **two-sided P values** or **observed significance levels** for the t statistics. The degrees of freedom used to calculate the P values is given by the Error DF from the ANOVA table. The P value for the independent variable tells us whether the independent variable has statistically signifiant predictive capability.

In theory, the P value for the constant could be used to determine whether the constant could be removed from the model. In practice, we do not usually do that. There are two reasons for this.

1. When there is no constant, the model is
$$Y = b_1 X ,$$
which forces Y to be 0 when X is 0. Even this is condition is appropriate (for example, no lean body mass means no strength), it is often wrong to place this constraint on the regression line. Most studies are performed with the independent variable far removed from 0. While a straight line may be appropriate for the range of data values studied, the relationship may not be a straight line all the way down to values of 0 for the predictor.
2. Standard practice (hierarchical modeling) is to include all simpler terms when a more complicated term is added to a model. Nothing is simpler than a constant. So if a change of Y with X is to be place in a model, the constant should be included, too. It could be argued this is a variant of (1).

The Analysis of Variance Table

The **Analysis of Variance** table is also known as the **ANOVA table** (for ANalysis Of VAriance). It tells the story of how the regression equation accounts for variablity in the response variable.

The column labeled **Source** has three rows: Regression, Residual, and Total. The column labeled **Sum of Squares** describes the variability in the response variable, Y.

The total amount of variability in the response is the **Total Sum of Squares**, $\sum (y_i - \bar{y})^2$. (The row labeled **Total** is sometimes labeled **Corrected Total**, where *corrected* refers to subtracting the sample mean before squaring and summing.) If a prediction had to be made without any other information, the best that could be done, in a certain sense, is to predict every value to be equal to the sample mean. The error--that is, the amount of variation in the data that can't be accounted for by this simple method--is given by the Total Sum of Squares.

When the regression model is used for prediction, the error (the amount of uncertainty that remains) is the variability about the regression line, $\sum (y_i - \hat{y}_i)^2$. This is the **Residual Sum of Squares** (*residual* for *left over*). It is sometimes called the Error Sum of Squares. The **Regression Sum of Squares** is the difference between the **Total Sum of Squares** and the Residual Sum of Squares. Since the **total sum of squares** is the total amount of variablity in the response and the **residual sum of squares** that still cannot be accounted for after the regression model is fitted, the **regression sum of squares** is the amount of variablity in the response that is accouned for by the regression model.

Each sum of squares has a corresponding degrees of freedom (DF) associated with it. Total df is *n-1*, one less than the number of observations. The Regression df is the number of independent variables in the model. For simple linear regression, the Regression df is 1. The Error df is the difference between the Total df and the Regression df. For simple linear regression, the residual df is *n-2*.

The **Mean Squares** are the Sums of Squares divided by the corresponding degrees of freedom.

The **F** statistic, also known as the **F ratio**, will be described in detail during the discussion of multiple regression. When there is only one predictor, the F statistic will be the square of the predictor variable's t statistic.

R^2 is the squared multiple correlation coefficient. It is also called the **Coefficient of Determination**. R^2 is the Regression sum of squares divided by the Total sum of squares, RegSS/TotSS. It is the fraction of the variability in the response that is fitted by the model. Since the Total SS is the sum of the Regression and Residual Sums of squares, R^2 can be rewritten as (TotSS-ResSS)/TotSS = 1- ResSS/TotSS. Some call R^2 *the proportion of the variance explained by the model*. I don't like the use of the word *explained* because it implies causality. However, the phrase is firmly entrenched in the literature. Even Fisher used it. If a model has perfect predictability, the Residual Sum of Squares will be 0 and $R^2=1$. If a model has no predictive capability, $R^2=0$. In practice, R^2 is never observed to be exactly 0 the same way the difference between the means of two samples drawn from the same population is never exaxctly 0 or a sample correlation coefficient is never exactly 0.

R, the multiple correlation coefficient and square root of R^2, is the correlation between the predicted and observed values. In simple linear regression, R will be equal to the magnitude correlation coefficient between X and Y. This is because the predicted values are b_0+b_1X. Neither multiplying by b_1 or adding b_0 affects the magnitude of the correlation coefficient. Therefore, the correlation between X and Y will be equal to the correlation between b_0+b_1X and Y, except for their sign if b_1 is negative.

Adjusted-R^2 will be described during the discussion of multiple regression.

The **Standard Error of the Estimate** (also known as the **Root Mean Square Error**) is the square root of the Residual Mean Square. It is the standard deviation of the data about the regression line, rather than about the sample mean. That is, it is

$$\sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} \text{ rather than } \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}$$

Last modified: 08/10/2007 22:33:52.