Standardized Coefficients

Task. How do you decide which of the Xs are most important for determining Y? In this handout, we discuss one possible (and controversial) answer to this question - the standardized regression coefficients.

Formulas. First, we will give the formulas and then explain their rationale:

General Case:	$b'_k = b_k * \frac{s_{x_k}}{s_y}$	As this formula shows, it is very easy to go from the metric to the standardized coefficients. There is no need to actually compute the standardized variables and run a new regression.
Two IV case:	$b'_1 = \frac{r_{y1} - (r_{12} * r_{y2})}{1 - r_{12}^2}$ $b'_2 = \frac{r_{y2} - (r_{12} * r_{y1})}{1 - r_{12}^2}$	Compare this to the formula for the metric coefficients. Note that correlations take the place of the corresponding variances and covariances.
1 IV case	$b' = r_{yx}$	In the one IV case, the standardized coefficient simply equals the correlation between Y and X

Rationale. The parameters a, b_1, b_2, etc., are often referred to as the metric regression coefficients. It is often difficult to say which of the X variables is most important in determining the value of the dependent variable, since the value of the regression coefficients depends on the choice of units to measure X. In the present example, this is not so problematic, since both education and job experience are measured in years. But suppose instead that our independent variables were education and IQ - how would we determine which variable was more important? The values of the metric coefficients would tell us little, since IQ and education are measured in very different ways.

For example, suppose the metric coefficient for education was 2.0, and the metric coefficient for IQ was 1.0. This would mean that each additional year of education was worth $2000 on average, and each 1-point increase in IQ was worth $1000 - but we certainly could not infer from this that education was more important than IQ in determining earnings. Keep in mind, too, that IQ scores are typically scaled to have a mean of 100 and a standard deviation of 16. This is an arbitrary scaling, however; they could just as easily divide all IQ scores by 2, giving a mean of 50 and an s.d. of 8. Such an arbitrary rescaling would change the value of the metric coefficient for IQ; instead of equaling 1, the coefficient would equal 2.

One proposed solution (much less popular than it used to be) has been to estimate regression models using "standardized" variables which are "metric-free." This is done by computing Z scores for each of the dependent and independent variables. That is,

$Y' = (Y - \hat{\mu}_Y)/s_y$, $X_1' = (X_1 - \hat{\mu}_{X_1})/s_1$, $X_2' = (X_2 - \hat{\mu}_{X_2})/s_2$, etc.

Conversely,

$Y = \hat{\mu}_Y + s_y Y'$, $X1 = \hat{\mu}_{X_1} + s_1 X1'$, $X2 = \hat{\mu}_{X_2} + s_2 X2'$

Each "standardized" variable has a mean of 0 and a variance of 1. Hence, for example, if Y' = 0, $Y = \hat{\mu}_Y = 24.415$. If Y' = 2, that means the individual has a score that is 2 standard deviation above the mean for Y; that is, $Y = \hat{\mu}_Y + s_y * 2 = 24.415 + 9.79 * 2 = 43.995$. For the first case in the present data set, Y = 5 ==> Y' = (5 - 24.415)/9.79 = -1.98. For the last case, Y = 48.3 ==> Y' = (48.3 - 24.415)/9.79 = 2.44.

Using the standardized variables, we estimate the model

$Y' = b_1'X1' + b_2'X2' + e'$

where b_1' and b_2' are the standardized regression coefficients. Note that we do not include the term a'. This is because $a' = \hat{\mu}_{Y'} - b_1' \hat{\mu}_{X_1'} - b_2' \hat{\mu}_{X_2'} = 0 - 0 - 0 = 0$.

Interpretation. We interpret the coefficients by saying that an increase of s_1 in X1 (i.e. 1 standard deviation) results, on average, in an increase of $b_1' * s_y$ in Y. For example, as we will see momentarily, $b_1' = .884$. Hence, increasing X1 by 4.48 (the standard deviation of X1) increases X1' by 1, which increases Y' (on average) by .884, or, equivalently, increases Y by $.884 * 9.79 = 8.65$. (You can confirm this by noting $b_1 = 1.933$, and $1.933 * 4.48 = 8.65$). Similarly, an increase of s_2 in X2 results in an average increase in Y of $b_2' * s_y$.

Hence, standardized coefficients tell you how increases in the independent variables affect relative position within the group. You can determine whether a 1 standard deviation change in one independent variable produces more of a change in relative position than a 1 standard deviation change in another independent variable.

Computation. We could actually compute the standardized variables, and then repeat steps a and b from the Multiple Regression handout. Given that we have made it this far, however, it is probably easier to note that

$$b_k' = b_k * \frac{s_{x_k}}{s_y}$$

Proof (Optional)	
Step	Rationale
$Y - \bar{y} = a + b_1X_1 + b_2X_2 + e - \bar{y}$	Subtract $\bar{Y}$ from both sides
$= \bar{y} - b_1 \bar{X}_1 - b_2 \bar{X}_2 + b_1X_1 + b_2X_2 + e - \bar{y}$	Substitute for a
$= b_1(X_1 - \bar{X}_1) + b_2(X_2 - \bar{X}_2) + e$	Rearrange terms
$= b_1 * s_1 * (X_1 - \bar{X}_1)/s_1 + b_2 * s_2 * (X_2 - \bar{X}_2)/s_2 + e$	Multiply and divide by s.d.s
$= b_1 * s_1 * X_1' + b_2 * s_2 * X_2' + e$	Substitute standardized X's
==> $(Y - \bar{y})/s_y = Y'$ $= b_1 * s_1/s_y * X_1' + b_2 * s_2/s_y * X_2' + e/s_y$	Divide both sides by s_y
$= b_1'X_1' + b_2'X_2' + e'$	Substitute standardized coefficients
==> $b_k' = b_k * s_k/s_y$	Q.E.D.

Hence, for this problem,

$b_1' = b_1 * s_1/s_y = 1.933 * 4.48 / 9.79 = .884$
$b_2' = b_2 * s_2/s_y = 0.649 * 5.46 / 9.79 = .362.$

Also, it easily follows that, if

H = the set of all the X (independent) variables,
G_k = the set of all the X variables *except* X_k, then,

$$s_{b_k'} = s_{b_k} * \frac{s_{x_k}}{s_y} = \sqrt{\frac{1 - R^2_{YH}}{(1 - R^2_{X_kG_k}) * (N - K - 1)}}$$

Ergo,

$s_{b1}' = s_{b1} * s_1/s_y = .210 * 4.48/9.79 = .096,$
$s_{b2}' = s_{b2} * s_2/s_y = .172 * 5.46/9.79 = .096.$ Or, equivalently,

$s_{b1}' = \sqrt{[(1 - R_{y12}^2)/((1 - R_{12}^2) * (N - K - 1))]}$

$= \sqrt{[(1 - .845)/((1 - .107^2) * 17)]} = .096$

(Note that, when there are only 2 independent variables, their standardized standard errors will be the same. This will generally not be true when there are more than 2 independent variables.)

Alternative computation (2 IV Case only!). Recall that, when there are two independent variables,

$b_1 = (s_2^2 * s_{y1} - s_{12} * s_{y2}) / (s_1^2 * s_2^2 - s_{12}^2)$
$b_2 = (s_1^2 * s_{y2} - s_{12} * s_{y1}) / (s_2^2 * s_1^2 - s_{12}^2)$

When variables are in standardized form, the correlation matrix is the same as the covariance matrix. That is, the variances of the standardized variables = 1, and the covariances equal the correlations. Hence, when there are <u>two</u> independent variables, you could also compute

$b_1' = (r_{y1} - r_{12} * r_{y2}) / (1 - r^2_{12}) =$
$(.845 + .107 * .268) / (1 - (-.107)^2 =$
$.874 / .989 = .884$

$b_2' = (r_{y2} - r_{12} * r_{y1}) / (1 - r^2_{12}) =$
$(.268 + .107 * .845) / (1 - (-.107)^2 =$
$.358 / .989 = .362$

(Recall too that, in the bivariate case, $b = s_{xy}/s_x^2$. Hence, when there is only one independent variable, $b' = r_{xy}$.)

[Optional] Other Analyses with Standardized Variables. Further, if you were so inclined, you could go through all the other steps outlined in our initial discussion of multiple regression. Among the things you would discover are SST = (n - 1), MST = 1, SSR = R^2 * (N - 1), MSR = R^2 * (N - 1)/K, and the values of the computed t's and F's are unaffected by the standardization. In practice, I don't think there would be much reason for wanting to do this.

Also, if you were presented with the results of an analysis done with standardized variables, and if you knew the s.d.'s of the unstandardized variables, it would be a fairly straightforward matter to compute the results of the analysis for the unstandardized variables. Just keep in mind that SST = s^2_y * SST' and SSE = s^2_y * SSE'. Also, SSR = R^2 * SST (regardless of whether variables are standardized or not).

Why might you want to do this? Possibly because results are only presented for the standardized variables, and you want to figure out what the unstandardized results are. (This is not an uncommon situation.) Also, computations are much simpler for standardized variables; depending on what you are interested in, it may be easier to work things out using the standardized variables and then convert back to the metric coefficients at the end. Hence, being able to convert standardized results back into metric results can occasionally be useful.

Going from standardized to metric. It is very easy to convert standardized coefficients back into metric coefficients, provided you know the standard deviations.

$$b_k = b_{k'} * \frac{s_y}{s_{x_k}}, \quad s_{b_k} = s_{b_{k'}} * \frac{s_y}{s_{x_k}}$$

For example,

$b_1 = b_1' * s_y/s_{x1} = .884 * 9.79 / 4.48 = 1.931,$
$b_2 = b_2' * s_y/s_{x2} = .362 * 9.79 / 5.46 = 0.649,$

$s_{b1} = s_{b1}' * s_y/s_{x1} = .096 * 9.79 / 4.48 = .210,$
$s_{b2} = s_{b2}' * s_y/s_{x2} = .096 * 9.79 / 5.46 = .172$

Computing R^2. Standardized coefficients provide an easy means for computing R^2.

$$R^2 = \sum b_{k'} r_{yk}; \text{ or,}$$

$$R^2 = b_{1'}^{\ 2} + b_{2'}^{\ 2} + 2 b_{1'} b_{2'} r_{12} \text{ (2 IV Case)}$$

Ergo,

$$R^2 = \sum b_{k'} r_{yk} = .884 * .845 + .362 * .268 = .844; \text{ or,}$$

$$R^2 = b_{1'}^{\ 2} + b_{2'}^{\ 2} + 2 b_{1'} b_{2'} r_{12} = .884^2 + .362^2 + 2 * .884 * .362 * -.107 = .844$$

Cautions about standardized coefficients:

✓ The coefficients can often be less intuitively meaningful

✓ The use of standardized coefficients can make it difficult to make comparisons across groups - because the standardization is different for each group.

For excellent discussions on standardized variables and coefficients, see Otis Dudley Duncan's book, **Structural Equation Modeling.** Also see

Kim, J. & G. Feree. 1981. "Standardization in Causal Analysis." **Sociological Methods and Research** 10(2):187-210.

We will discuss these issues much more in Stats II.