Standardized Regressions

Consider the usual regression equation with intercept,

$$Y = \alpha \mathbf{1} + X\beta + \epsilon \,, \tag{1}$$

where we have separated out the intercept. Let the OLS fit be

$$Y = \widehat{\alpha}\mathbf{1} + X\widehat{\beta} + e \,. \tag{2}$$

Now suppose we standardize all the variables. That is,

$$\widetilde{Y} = (Y - \bar{Y}\mathbf{1})/\mathrm{sd}(Y)$$

is the standardized Y, and $\widetilde{X}$ denotes the matrix whose columns are the standardized columns of X. Let $X^{[k]}$ denote the kth column of X, with mean m_k and sd s_k. Note that $X^{[k]} - m_k\mathbf{1}$ is the column where m_k has been subtracted from each entry of $X^{[k]}$. Thus,

$$\widetilde{X}^{[k]} = (X^{[k]} - m_k\mathbf{1})/s_k \,.$$

We can solve this for $X^{[k]}$ to get

$$X^{[k]} = s_k\widetilde{X}^{[k]} + m_k\mathbf{1} \,.$$

Therefore, substitution in (2) yields

$$Y = \widehat{\alpha}\mathbf{1} + X\widehat{\beta} + e = \widehat{\alpha}\mathbf{1} + \sum_{k=1}^{p} X^{[k]}\widehat{\beta}_k + e = \widehat{\alpha}\mathbf{1} + \sum_{k=1}^{p}(s_k\widetilde{X}^{[k]} + m_k\mathbf{1})\widehat{\beta}_k + e$$

$$= (\widehat{\alpha} + \sum_{k=1}^{p} m_k\widehat{\beta}_k)\mathbf{1} + \sum_{k=1}^{p} s_k\widetilde{X}^{[k]}\widehat{\beta}_k + e \,. \tag{3}$$

Using the definition of $\widetilde{Y}$, this gives

$$\widetilde{Y} = b\mathbf{1} + \widetilde{X}\gamma + e/\mathrm{sd}(Y) \tag{4}$$

for some real b and the vector γ with

$$\gamma_k = s_k\widehat{\beta}_k/\mathrm{sd}(Y) \,. \tag{5}$$

1

If D is the diagonal matrix whose (k, k)-entry is s_k, the standard deviation of the kth column of X, then we may write (5) as

$$\gamma = D\widehat{\beta}/\mathrm{sd}(Y)\,. \tag{6}$$

Now recall that $e \perp \mathbf{1}$. Since we standardized, we also have $\widetilde{Y} \perp \mathbf{1}$ and $\widetilde{X} \perp \mathbf{1}$. Therefore, taking the dot product of both sides of (4) with $\mathbf{1}$, we obtain $b = 0$. Thus, we may use a new residual $d := e/\mathrm{sd}(Y)$ to write (4) as

$$\widetilde{Y} = \widetilde{X}\gamma + d\,.$$

Furthermore, $d \perp \widetilde{X}$ since the columns of $\widetilde{X}$ are linear combinations of the columns of X and $\mathbf{1}$, and we know that $d \perp \mathbf{1}$ and $d \perp X$. **That is, OLS regression of $\widetilde{Y}$ on $\widetilde{X}$ gives γ.**

The first thing we conclude is that the "estimated standardized-regression coefficients" γ are obtained from the usual estimated regression coefficients $\widehat{\beta}$ by "changing units", i.e., $\gamma_k = \widehat{\beta}_k \mathrm{sd}(X^{[k]})/\mathrm{sd}(Y)$, which is equation (5). There are no parameters in a standardized regression since the standardization is defined in terms of the data. Thus, we aren't really estimating the non-existent parameters. (For another interpretation that involves standardizing Y at the population level, and which leads to a genuine regression with new parameters, see Freedman's note, "Comments on standardizing path diagrams: what are the parameters?" at `http://www.stat.berkeley.edu/~census/standard.pdf`. This is a transformation of (1) rather than of (2), and involves using $(Y - \alpha\mathbf{1})/\mathrm{SD}(Y_1)$ rather than $\widetilde{Y}$. It is more complicated than what we do and it introduces new bias.) However, we may regard γ as estimating β with a change of scale that depends on the unstandardized data. If we know the standard deviations of the unstandardized data, then we may calculate $\widehat{\beta}$ from γ. But if we know only standardized data, then although we can calculate γ, we cannot calculate $\widehat{\beta}$ since this change of scale will be unknown (unless we are told $\mathrm{sd}(X^{[k]})$ and $\mathrm{sd}(Y)$).

The next thing is to find the SEs for γ. Going from X to $\widetilde{X}$ involved two steps: first, we subtracted a multiple of $\mathbf{1}$ from each column to center it, getting, say, the matrix Z; second, we divided each column by its standard deviation. Notice that using a new design matrix obtained from subtracting multiples of $\mathbf{1}$ from the columns of X does not change β or $\widehat{\beta}$, though it does change the intercept. (Look at the equations (1) and (2), or else (3).) Let's write

$$Y = \widehat{a}\mathbf{1} + Z\widehat{\beta} + e$$

for the regression of Y on $\mathbf{1}$ and Z. Since $\mathbf{1} \perp Z$ by design of Z, we have that the new design matrix $[\mathbf{1}\ Z]$ satisfies $[\mathbf{1}\ Z]'[\mathbf{1}\ Z] = \begin{bmatrix} n & \mathbf{0} \\ \mathbf{0} & Z'Z \end{bmatrix}$; here, n is the length of Y, as usual. Therefore,

$$\operatorname{Cov}\left(\begin{bmatrix} \widehat{a} \\ \widehat{\beta} \end{bmatrix} \Big| X \right) = \operatorname{Cov}\left(\begin{bmatrix} \widehat{a} \\ \widehat{\beta} \end{bmatrix} \Big| Z \right) = \sigma^2 \left([\mathbf{1}\ Z]'[\mathbf{1}\ Z]\right)^{-1} = \sigma^2 \begin{bmatrix} 1/n & \mathbf{0} \\ \mathbf{0} & (Z'Z)^{-1} \end{bmatrix}.$$

In particular,

$$\operatorname{Cov}(\widehat{\beta} \mid X) = \operatorname{Cov}(\widehat{\beta} \mid Z) = \sigma^2 (Z'Z)^{-1}.$$

The second step in creating $\widetilde{X}$ from X involved dividing by the standard deviations: $\widetilde{X} = ZD^{-1}$. From this and (6), we deduce that

$$\operatorname{Cov}(\gamma \mid X) = \operatorname{Cov}\left(D\widehat{\beta}/\operatorname{sd}(Y) \mid X\right) = D\operatorname{Cov}(\widehat{\beta} \mid X)D/\operatorname{var}(Y) = \sigma^2 D(Z'Z)^{-1}D/\operatorname{var}(Y)$$
$$= \sigma^2 (D^{-1}Z'ZD^{-1})^{-1}/\operatorname{var}(Y) = \sigma^2 (\widetilde{X}'\widetilde{X})^{-1}/\operatorname{var}(Y).$$

We don't know σ; if we are given $\widetilde{Y}$ and not Y, then we don't know $\operatorname{var}(Y)$ either. But recall that $\widehat{\sigma}^2 = \|e\|^2/(n-p)$. We don't know e, but we do know $d = e/\operatorname{sd}(Y)$. Thus, we have

$$\widehat{\sigma}^2/\operatorname{var}(Y) = \|d\|^2/(n-p),$$

which gives us

$$\widehat{\operatorname{Cov}}(\gamma \mid \widetilde{X}) := \widehat{\sigma}^2 (\widetilde{X}'\widetilde{X})^{-1}/\operatorname{var}(Y) = \frac{\|d\|^2}{n-p}(\widetilde{X}'\widetilde{X})^{-1}. \tag{7}$$

This is just the formula you would use if you treated a standardized-regression equation as an ordinary regression equation for the purposes of OLS, except that p is one more than the number of columns of $\widetilde{X}$.

Consider the null hypothesis that $\beta_k = 0$. We test it using $t := \widehat{\beta}_k/\widehat{\operatorname{SE}}(\widehat{\beta}_k)$. By (5), $\gamma_k = \widehat{\beta}_k s_k/\operatorname{sd}(Y)$, whence $\operatorname{SE}(\gamma_k) = \operatorname{SE}(\widehat{\beta}_k)s_k/\operatorname{sd}(Y)$. Dividing the two preceding equations, we obtain $\widehat{\beta}_k/\operatorname{SE}(\widehat{\beta}_k) = \gamma_k/\operatorname{SE}(\gamma_k)$. Now by (7), we have

$$\widehat{\operatorname{SE}}(\gamma_k) = \frac{\|d\|}{\sqrt{n-p}}(\widetilde{X}'\widetilde{X})^{-1}_{k,k} = \frac{\widehat{\sigma}}{\operatorname{sd}(Y)}(\widetilde{X}'\widetilde{X})^{-1}_{k,k} = \frac{\widehat{\sigma}}{\sigma}\operatorname{SE}(\gamma_k).$$

Of course, we also have that $\widehat{\operatorname{SE}}(\widehat{\beta}_k) = (\widehat{\sigma}/\sigma)\operatorname{SE}(\widehat{\beta}_k)$. Therefore $t = \gamma_k/\widehat{\operatorname{SE}}(\gamma_k)$. ***That is, the t-test statistic for γ_k is the same as for $\widehat{\beta}_k$.***

In addition, since $d = e/\operatorname{sd}(Y)$ as a general rule, if we are testing a smaller model, we will also have $d^{(\mathrm{s})} = e^{(\mathrm{s})}/\operatorname{sd}(Y)$. It follows that

$$F = \frac{\left(\|e^{(\mathrm{s})}\|^2 - \|e\|^2\right)/p_0}{\|e\|^2/(n-p)} = \frac{\left(\|d^{(\mathrm{s})}\|^2 - \|d\|^2\right)/p_0}{\|d\|^2/(n-p)}.$$

Therefore, the F-test statistic for any subset of γ_k is the same as for the corresponding subset of $\widehat{\beta}_k$. Also, R^2 is unchanged for the standardized regression since

$$1 - R^2 = \frac{\mathrm{var}(e)}{\mathrm{var}(Y)} = \frac{\|e\|^2}{n\,\mathrm{var}(Y)} = \frac{\|d\|^2}{n} = \frac{\|d\|^2}{\|\widetilde{Y}\|^2}\,.$$

Note that we use the definition of R^2 with an intercept for the original regression equation and the definition of R^2 without an intercept for the standardized regression equation (though in the latter case, it wouldn't matter which definition we used).

The t-test statistic for a linear combination $c'\gamma$ of the γ_k is not the same as the t-statistic for $c'\beta$ in general because different β_k get divided by different quantities $\mathrm{sd}(\beta_k)$. Some other linear combination $f'\beta$ would have to be used, but it would be impossible to know which f to use if the $\mathrm{sd}(\beta_k)$ are unknown. Nevertheless, linear combinations may seem to make more sense for γ than for $\widehat{\beta}$, as in Section 6.3, since "effect sizes" are comparable. However, this is an illusion because γ does not estimate parameters (recall there is no such thing as a standardized parameter). If the data gathered changed $\mathrm{sd}(X^{[k]})$, then γ_k would change as well, but there might not be any real difference in effect. (Thus, even with Freedman's interpretation of standardized regression, where there are real parameters, or even if we standardize only on the right-hand side and not Y, we can't get a stable interpretation of γ. Nevertheless, if you think of standardizing by sample quantities as approximately standardizing by population quantities, then γ does approximate actual parameters and therefore testing whether $c'\gamma = 0$ will make approximate sense—if indeed the sample is a good approximation of the population.)

If we don't have data or even standardized data, but have only correlations (as in Section 6.1), we can still compute $\gamma = (\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'\widetilde{Y}$ since we need only $\widetilde{X}'\widetilde{Y}$ and $\widetilde{X}'\widetilde{X}$, which are formed from the correlations: the kth entry of $\widetilde{X}'\widetilde{Y}$ is n times the correlation between $X^{[k]}$ and $\widetilde{Y}$, while the (j,k)-entry of $\widetilde{X}'\widetilde{X}$ equals n times the correlation between $X^{[j]}$ and $X^{[k]}$. Similarly, we can get $\|d\|^2$, which we need to estimate errors, as follows:

$$\|d\|^2 = d'd = d'(\widetilde{Y} - \widetilde{X}\gamma) = d'\widetilde{Y} = (\widetilde{Y} - \widetilde{X}\gamma)'\widetilde{Y} = n - \gamma'\widetilde{X}'\widetilde{Y}\,.$$

Here, we used the fact that $d \perp \widetilde{X}$. Thus, after we calculate $\widetilde{X}'\widetilde{Y}$ and then γ, we may calculate $\|d\|^2$.