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. Then

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

so

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

Therefore,

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

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

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

for some real b and the vector γ with $\gamma_k = s_k\widehat{\beta}_k/\mathrm{sd}(Y)$. If D is the diagonal matrix whose (k,k)-entry is the standard deviation of the kth column of X, then we may write $\gamma = D\widehat{\beta}/\mathrm{sd}(Y)$. 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, $b = 0$. Thus, we may use a new residual $d = e/\mathrm{sd}(Y)$ to write

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

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 by "changing units", i.e., $\gamma_k = \widehat{\beta}_k \mathrm{sd}(X^{(k)})/\mathrm{sd}(Y)$. 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. 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 $\widehat{\beta}$. But this change of scale may be unknown if we know only standardized data.

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).) Let's write

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

for the regression of Y on $\mathbf{1}$ and Z. Since $\mathbf{1} \perp Z$, 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,

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

In particular,

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

Since $\widetilde{X} = ZD^{-1}$, we deduce that

$$\begin{aligned}\mathrm{Cov}(\gamma \mid X) &= \mathrm{Cov}(D\widehat{\beta}/\mathrm{sd}(Y) \mid X) = D\,\mathrm{Cov}(\widehat{\beta} \mid X)D/\mathrm{var}(Y) = \sigma^2 D(Z'Z)^{-1}D/\mathrm{var}(Y) \\ &= \sigma^2(D^{-1}Z'ZD^{-1})^{-1}/\mathrm{var}(Y) = \sigma^2(\widetilde{X}'\widetilde{X})^{-1}/\mathrm{var}(Y).\end{aligned}$$

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

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

which gives us

$$\widehat{\mathrm{Cov}}(\gamma \mid \widetilde{X}) = \frac{\|d\|^2}{n-p}(\widetilde{X}'\widetilde{X})^{-1}\,.$$

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 $\widetilde{X}$. Since for each k, we have that γ_k is a multiple of $\widehat{\beta}_k$, this means that the t-test statistic for γ_k is the same as for $\widehat{\beta}_k$. In addition, since $d = e/\mathrm{sd}(Y)$, if we are testing a smaller model, we will also have $d^{(\mathrm{s})} = e^{(\mathrm{s})}/\mathrm{sd}(Y)$, whence ***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 statistics for linear combinations of γ_k are the same only if you change the combination by the change of units, which will be impossible if the units are unknown. But linear combinations may make more sense for γ than for $\widehat{\beta}$ since "effect sizes" are comparable.

If we don't have data or even standardized data, but have only correlations (as in Section 6.1), we can still compute γ since we need only $\widetilde{X}'\widetilde{Y}$ and $\widetilde{X}'\widetilde{X}$, which are formed from the correlations. Similarly, we can get $\|d\|^2$, which we need to estimate errors, as follows:

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

Here, we used the fact that $d \perp \widetilde{X}$. We now know γ and have already calculated $\widetilde{X}'\widetilde{Y}$.